An Introduction to MULTIVARIATE STATISTICAL ANALYSIS

2nd Edition

T.W. Anderson

A Volume in the Wiley Series in Probability and Mathematical Statistics:
Vic Barnett, Ralph A. Bradley, J. Stuart Hunter, David G. Kendall,
Rupert G. Miller, Jr., Stephen M. Stigler, Geoffrey S. Watson
Advisory Editors

An Introduction to Multivariate Statistical Analysis

Second Edition

T. W. ANDERSON
Professor of Statistics and Economics
Stanford University

JOHN WILEY & SONS
New York Chichester Brisbane Toronto Singapore

***Library of Congress Cataloging in Publication Data*:**

Anderson, T. W. (Theodore Wilbur), 1918–
An introduction to multivariate statistical analysis.

Bibliography: p.
Includes index.
1. Multivariate analysis. I. Title.

QA278.A516 1984 519.5′35 84-7334
ISBN 0-471-88987-3

Printed in the United States of America

10 9 8 7 6 5 4 3 2

Preface to the Second Edition

Twenty-six years have passed since the first edition of this book was published. During that time great advances have been made in multivariate statistical analysis—particularly in the areas treated in that volume. This new edition purports to bring the original edition up to date by substantial revision, rewriting, and additions. The basic approach has been maintained, namely, a mathematically rigorous development of statistical methods for observations consisting of several measurements or characteristics of each subject and a study of their properties. The general outline of topics has been retained.

The method of maximum likelihood has been augmented by other considerations. In point estimation of the mean vector and covariance matrix alternatives to the maximum likelihood estimators that are better with respect to certain loss functions, such as Stein and Bayes estimators, have been introduced. In testing hypotheses likelihood ratio tests have been supplemented by other invariant procedures. New results on distributions and asymptotic distributions are given; some significance points are tabulated. Properties of these procedures, such as power functions, admissibility, unbiasedness, and monotonicity of power functions, are studied. Simultaneous confidence intervals for means and covariances are developed. A chapter on factor analysis replaces the chapter sketching miscellaneous results in the first edition. Some new topics, including simultaneous equations models and linear functional relationships, are introduced. Additional problems present further results.

It is impossible to cover all relevant material in this book; what seems most important has been included. For a comprehensive listing of papers until 1966 and books until 1970 the reader is referred to *A Bibliography of Multivariate Statistical Analysis* by Anderson, Das Gupta, and Styan (1972). Further references can be found in *Multivariate Analysis: A Selected and Abstracted Bibliography, 1957–1972* by Subrahmaniam and Subrahmaniam (1973).

I am in debt to many students, colleagues, and friends for their suggestions and assistance; they include Yasuo Amemiya, James Berger, Byoung-Seon Choi, Arthur Cohen, Margery Cruise, Somesh Das Gupta, Kai-Tai Fang, Gene Golub, Aaron Han, Takeshi Hayakawa, Jogi Henna, Huang Hsu, Fred Huffer, Mituaki Huzii, Jack Kiefer, Mark Knowles, Sue Leurgans, Alex McMillan, Masashi No, Ingram Olkin, Kartik Patel, Michael Perlman, Allen Sampson, Ashis Sen Gupta, Andrew Siegel, Charles Stein, Patrick Strout, Akimichi Takemura, Joe Verducci, Marlos Viana, and Y. Yajima. I was helped in preparing the manuscript by Dorothy Anderson, Alice Lundin, Amy Schwartz, and Pat Struse. Special thanks go to Johanne Thiffault and George P. H. Styan for their precise attention. Support was contributed by the Army Research Office, the National Science Foundation, the Office of Naval Research, and IBM Systems Research Institute.

Seven tables of significance points are given in Appendix B to facilitate carrying out test procedures. Tables 1, 5, and 7 are Tables 47, 50, and 53, respectively, of *Biometrika Tables for Statisticians*, Vol. 2, by E. S. Pearson and H. O. Hartley; permission of the Biometrika Trustees is hereby acknowledged. Table 2 is made up from three tables prepared by A. W. Davis and published in *Biometrika* (1970a), *Annals of the Institute of Statistical Mathematics* (1970b), and *Communications in Statistics, B. Simulation and Computation* (1980). Tables 3 and 4 are Tables 6.3 and 6.4, respectively, of *Concise Statistical Tables*, edited by Ziro Yamauti (1977) and published by the Japanese Standards Association; this book is a concise version of *Statistical Tables and Formulas with Computer Applications, JSA-1972*. Table 6 is Table 3 of *The Distribution of the Sphericity Test Criterion*, ARL 72-0154, by B. N. Nagarsenker and K. C. S. Pillai, Aerospace Research Laboratories (1972). The author is indebted to the authors and publishers listed above for permission to reproduce these tables.

T. W. ANDERSON

Stanford, California
June 1984

Preface to the First Edition

This book has been designed primarily as a text for a two-semester course in multivariate statistics. It is hoped that the book will also serve as an introduction to many topics in this area to statisticians who are not students and will be used as a reference by other statisticians.

For several years the book in the form of dittoed notes has been used in a two-semester sequence of graduate courses at Columbia University; the first six chapters constituted the text for the first semester, emphasizing correlation theory. It is assumed that the reader is familiar with the usual theory of univariate statistics, particularly methods based on the univariate normal distribution. A knowledge of matrix algebra is also a prerequisite; however, an appendix on this topic has been included.

It is hoped that the more basic and important topics are treated here, though to some extent the coverage is a matter of taste. Some of the more recent and advanced developments are only briefly touched on in the last chapter.

The method of maximum likelihood is used to a large extent. This leads to reasonable procedures; in some cases it can be proved that they are optimal. In many situations, however, the theory of desirable or optimum procedures is lacking.

Over the years this manuscript has been developed, a number of students and colleagues have been of considerable assistance. Allan Birnbaum, Harold Hotelling, Jacob Horowitz, Howard Levene, Ingram Olkin, Gobind Seth, Charles Stein, and Henry Teicher are to be mentioned particularly. Acknowledgments are also due to other members of the Graduate Mathematical

Statistics Society at Columbia University for aid in the preparation of the manuscript in dittoed form. The preparation of this manuscript was supported in part by the Office of Naval Research.

T. W. ANDERSON

Center for Advanced Study
in the Behavioral Sciences
Stanford, California
December 1957

Contents

CHAPTER 1

Introduction

1.1. MULTIVARIATE STATISTICAL ANALYSIS

Multivariate statistical analysis is concerned with data that consist of sets of measurements on a number of individuals or objects. The sample data may be heights and weights of some individuals drawn randomly from a population of schoolchildren in a given city, or the statistical treatment may be made on a collection of measurements, such as lengths and widths of petals and lengths and widths of sepals of iris plants taken from two species, or one may study the scores on batteries of mental tests administered to a number of students.

The measurements made on a single individual can be assembled into a column vector. We think of the entire vector as an observation from a multivariate population or distribution. When the individual is drawn randomly, we consider the vector as a random vector with a distribution or probability law describing that population. The set of observations on all individuals in a sample constitute a sample of vectors, and the vectors set side by side make up the matrix of observations.[†] The data to be analyzed then are thought of as displayed in a matrix or in several matrices.

We shall see that it is helpful in visualizing the data and understanding the methods to think of each observation vector as constituting a point in a Euclidean space, each coordinate corresponding to a measurement or variable. Indeed, an early step in the statistical analysis is plotting the data; since most statisticians are limited to two-dimensional plots, two coordinates of the observation are plotted in turn.

[†]When data are listed on paper by individual, it is natural to print the measurements on one individual as a row of the table; then one individual corresponds to a *row* vector. Since we prefer to operate algebraically with column vectors, we have chosen to treat observations in terms of *column* vectors. (In practice, the basic data set may well be on cards, tapes, or disks.)

Characteristics of a univariate distribution of essential interest are the mean as a measure of location and the standard deviation as a measure of variability; similarly the mean and standard deviation of a univariate sample are important summary measures. In multivariate analysis, the means and variances of the separate measurements—for distributions and for samples—have corresponding relevance. An essential aspect, however, of multivariate analysis is the dependence between the different variables. The dependence between two variables may involve the covariance between them, that is, the average products of their deviations from their respective means. The covariance standardized by the corresponding standard deviations is the correlation coefficient; it serves as a measure of degree of dependence. A set of summary statistics is the mean vector (consisting of the univariate means) and the covariance matrix (consisting of the univariate variances and bivariate covariances). An alternative set of summary statistics with the same information is the mean vector, the set of standard deviations, and the correlation matrix. Similar parameter quantities describe location, variability, and dependence in the population or for a probability distribution. The multivariate *normal* distribution is completely determined by its mean vector and covariance matrix, and the sample mean vector and covariance matrix constitute a sufficient set of statistics.

The measurement and analysis of dependence between variables, between sets of variables, and between variables and sets of variables, are fundamental to multivariate analysis. The multiple correlation coefficient is an extension of the notion of correlation to the relationship of one variable to a set of variables. The partial correlation coefficient is a measure of dependence between two variables when the effects of other correlated variables have been removed. The various correlation coefficients computed from samples are used to estimate corresponding correlation coefficients of distributions. In this book tests of hypotheses of independence are developed. The properties of the estimators and test procedures are studied for sampling from the multivariate normal distribution.

A number of statistical problems arising in multivariate populations are straightforward analogs of problems arising in univariate populations; the suitable methods for handling these problems are similarly related. For example, in the univariate case we may wish to test the hypothesis that the mean of a variable is zero; in the multivariate case we may wish to test the hypothesis that the vector of the means of several variables is the zero vector. The analog of the Student t-test for the first hypothesis is the generalized T^2-test. The analysis of variance of a single variable is adapted to vector observations; in regression analysis, the dependent quantity may be a vector variable. A comparison of variances is generalized into a comparison of covariance matrices.

The test procedures of univariate statistics are generalized to the multivariate case in such ways that the dependence between variables is taken into account. These methods may not depend on the coordinate system; in other terms, the procedures are invariant with respect to linear transformations that leave the null hypothesis invariant. In some problems there may be families of tests that are invariant; then choices must be made. Optimal properties of the tests are considered.

For some other purposes, however, it may be important to select a coordinate system so that the variates have desired statistical properties. One might say that they involve characterizations of inherent properties of normal distributions and of samples. These are closely related to the algebraic problems of canonical forms of matrices. An example is finding the normalized linear combination of variables with maximum or minimum variance (finding principal components); this amounts to finding a rotation of axes that carries the covariance matrix to diagonal form. Another example is characterizing the dependence between two sets of variates (finding canonical correlations). These problems involve the characteristic roots and vectors of various matrices. The statistical properties of the corresponding sample quantities are treated.

Some statistical problems arise in models in which means and covariances are restricted. Factor analysis may be based on a model with a (population) covariance matrix that is the sum of a positive definite diagonal matrix and a positive semidefinite matrix of low rank; linear structural relationships may have a similar formulation. The simultaneous equations system of econometrics is another example of a special model.

1.2. THE MULTIVARIATE NORMAL DISTRIBUTION

The statistical methods treated in this book can be developed and evaluated in the context of the multivariate normal distribution, though many of the procedures are useful and effective when the distribution sampled is not normal. A major reason for basing statistical analysis on the normal distribution is that this probabilistic model approximates well the distribution of continuous measurements in many sampled populations. In fact, most of the