APPLIED MULTIVARIATE STATISTICAL ANALYSIS

SIXTH EDITION

RICHARD A. JOHNSON

DEAN W. WICHERN

JOHNSON
WICHERN

APPLIED MULTIVARIATE STATISTICAL ANALYSIS

SIXTH
EDITION

PEARSON

Universidad Carlos III de Madrid
S
519.237

StudentAid.ed.gov
FUNDING YOUR FUTURE

PEARSON
Prentice
Hall

Upper Saddle River, NJ 07458
www.prenhall.com

ISBN-13: 978-0-13-187715-3
ISBN-10 0-13-187715-1
EAN
9 780131 877153
90000

Applied Multivariate Statistical Analysis

SIXTH EDITION

Applied Multivariate Statistical Analysis

RICHARD A. JOHNSON
University of Wisconsin—Madison

DEAN W. WICHERN
Texas A&M University

CC SS JJ

PEARSON
Prentice Hall

Upper Saddle River, New Jersey 07458

BIBLIOTECA
0000417504

brary of Congress Cataloging-in-Publication Data
hnson, Richard A.
Statistical analysis/Richard A. Johnson.—6th ed.
Dean W. Winchern
p. cm.
Includes index.
ISBN 0-13-187715-1
1. Statistical Analysis

IP Data Available

xecutive Acquisitions Editor: *Petra Recter*
Vice President and Editorial Director, Mathematics: *Christine Hoag*
roject Manager: *Michael Bell*
roduction Editor: *Debbie Ryan*
enior Managing Editor: *Linda Mihatov Behrens*
anufacturing Buyer: *Maura Zaldivar*
Associate Director of Operations: *Alexis Heydt-Long*
arketing Manager: *Wayne Parkins*
Marketing Assistant: *Jennifer de Leeuwerk*
Editorial Assistant/Print Supplements Editor: *Joanne Wendelken*
rt Director: *Jayne Conte*
Director of Creative Service: *Paul Belfanti*
over Designer: *Bruce Kenselaar*
rt Studio: *Laserswords*

© 2007 Pearson Education, Inc.
Pearson Prentice Hall
Pearson Education, Inc.
Upper Saddle River, NJ 07458

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Pearson Prentice Hall™ is a trademark of Pearson Education, Inc.

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN-13: 978-0-13-187715-3
ISBN-10: 0-13-187715-1

Pearson Education LTD., *London*
Pearson Education Australia PTY, Limited, *Sydney*
Pearson Education Singapore, Pte. Ltd
Pearson Education North Asia Ltd, *Hong Kong*
Pearson Education Canada, Ltd., *Toronto*
Pearson Educación de Mexico, S.A. de C.V.
Pearson Education–Japan, *Tokyo*
Pearson Education Malaysia, Pte. Ltd

To the memory of my mother and my father.
R. A. J.

To Dorothy, Michael, and Andrew.
D. W. W.

Contents

PREFACE xv

1 ASPECTS OF MULTIVARIATE ANALYSIS 1

1.1 Introduction 1
1.2 Applications of Multivariate Techniques 3
1.3 The Organization of Data 5
Arrays, 5
Descriptive Statistics, 6
Graphical Techniques, 11
1.4 Data Displays and Pictorial Representations 19
Linking Multiple Two-Dimensional Scatter Plots, 20
Graphs of Growth Curves, 24
Stars, 26
Chernoff Faces, 27
1.5 Distance 30
1.6 Final Comments 37
Exercises 37
References 47

2 MATRIX ALGEBRA AND RANDOM VECTORS 49

2.1 Introduction 49
2.2 Some Basics of Matrix and Vector Algebra 49
Vectors, 49
Matrices, 54
2.3 Positive Definite Matrices 60
2.4 A Square-Root Matrix 65
2.5 Random Vectors and Matrices 66
2.6 Mean Vectors and Covariance Matrices 68
Partitioning the Covariance Matrix, 73
The Mean Vector and Covariance Matrix for Linear Combinations of Random Variables, 75
Partitioning the Sample Mean Vector and Covariance Matrix, 77
2.7 Matrix Inequalities and Maximization 78

Supplement 2A: Vectors and Matrices: Basic Concepts 82
Vectors, 82
Matrices, 87
Exercises 103
References 110

3 SAMPLE GEOMETRY AND RANDOM SAMPLING 111

3.1 Introduction 111
3.2 The Geometry of the Sample 111
3.3 Random Samples and the Expected Values of the Sample Mean and Covariance Matrix 119
3.4 Generalized Variance 123
Situations in which the Generalized Sample Variance Is Zero, 129
Generalized Variance Determined by $|R|$ and Its Geometrical Interpretation, 134
Another Generalization of Variance, 137
3.5 Sample Mean, Covariance, and Correlation As Matrix Operations 137
3.6 Sample Values of Linear Combinations of Variables 140
Exercises 144
References 148

4 THE MULTIVARIATE NORMAL DISTRIBUTION 149

4.1 Introduction 149
4.2 The Multivariate Normal Density and Its Properties 149
Additional Properties of the Multivariate Normal Distribution, 156
4.3 Sampling from a Multivariate Normal Distribution and Maximum Likelihood Estimation 168
The Multivariate Normal Likelihood, 168
Maximum Likelihood Estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, 170
Sufficient Statistics, 173
4.4 The Sampling Distribution of $\overline{\mathbf{X}}$ and $\mathbf{S}$ 173
Properties of the Wishart Distribution, 174
4.5 Large-Sample Behavior of $\overline{\mathbf{X}}$ and $\mathbf{S}$ 175
4.6 Assessing the Assumption of Normality 177
Evaluating the Normality of the Univariate Marginal Distributions, 177
Evaluating Bivariate Normality, 182
4.7 Detecting Outliers and Cleaning Data 187
Steps for Detecting Outliers, 189
4.8 Transformations to Near Normality 192
Transforming Multivariate Observations, 195
Exercises 200
References 208

5 INFERENCES ABOUT A MEAN VECTOR 210

5.1 Introduction 210
5.2 The Plausibility of $\boldsymbol{\mu}_0$ as a Value for a Normal Population Mean 210
5.3 Hotelling's T^2 and Likelihood Ratio Tests 216
General Likelihood Ratio Method, 219
5.4 Confidence Regions and Simultaneous Comparisons of Component Means 220
Simultaneous Confidence Statements, 223
A Comparison of Simultaneous Confidence Intervals with One-at-a-Time Intervals, 229
The Bonferroni Method of Multiple Comparisons, 232
5.5 Large Sample Inferences about a Population Mean Vector 234
5.6 Multivariate Quality Control Charts 239
Charts for Monitoring a Sample of Individual Multivariate Observations for Stability, 241
Control Regions for Future Individual Observations, 247
Control Ellipse for Future Observations, 248
T^2-Chart for Future Observations, 248
Control Charts Based on Subsample Means, 249
Control Regions for Future Subsample Observations, 251
5.7 Inferences about Mean Vectors when Some Observations Are Missing 251
5.8 Difficulties Due to Time Dependence in Multivariate Observations 256
Supplement 5A: Simultaneous Confidence Intervals and Ellipses as Shadows of the p-Dimensional Ellipsoids 258
Exercises 261
References 272

6 COMPARISONS OF SEVERAL MULTIVARIATE MEANS 273

6.1 Introduction 273
6.2 Paired Comparisons and a Repeated Measures Design 273
Paired Comparisons, 273
A Repeated Measures Design for Comparing Treatments, 279
6.3 Comparing Mean Vectors from Two Populations 284
Assumptions Concerning the Structure of the Data, 284
Further Assumptions When n_1 and n_2 Are Small, 285
Simultaneous Confidence Intervals, 288
The Two-Sample Situation When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, 291
An Approximation to the Distribution of T^2 for Normal Populations When Sample Sizes Are Not Large, 294
6.4 Comparing Several Multivariate Population Means (One-Way Manova) 296
Assumptions about the Structure of the Data for One-Way MANOVA, 296

A Summary of Univariate ANOVA, 297
Multivariate Analysis of Variance (MANOVA), 301
6.5 Simultaneous Confidence Intervals for Treatment Effects 308
6.6 Testing for Equality of Covariance Matrices 310
6.7 Two-Way Multivariate Analysis of Variance 312
Univariate Two-Way Fixed-Effects Model with Interaction, 312
Multivariate Two-Way Fixed-Effects Model with Interaction, 315
6.8 Profile Analysis 323
6.9 Repeated Measures Designs and Growth Curves 328
6.10 Perspectives and a Strategy for Analyzing Multivariate Models 332
Exercises 337
References 358

7 *MULTIVARIATE LINEAR REGRESSION MODELS* 360

7.1 Introduction 360
7.2 The Classical Linear Regression Model 360
7.3 Least Squares Estimation 364
Sum-of-Squares Decomposition, 366
Geometry of Least Squares, 367
Sampling Properties of Classical Least Squares Estimators, 369
7.4 Inferences About the Regression Model 370
Inferences Concerning the Regression Parameters, 370
Likelihood Ratio Tests for the Regression Parameters, 374
7.5 Inferences from the Estimated Regression Function 378
Estimating the Regression Function at $\mathbf{z}_0$, 378
Forecasting a New Observation at $\mathbf{z}_0$, 379
7.6 Model Checking and Other Aspects of Regression 381
Does the Model Fit?, 381
Leverage and Influence, 384
Additional Problems in Linear Regression, 384
7.7 Multivariate Multiple Regression 387
Likelihood Ratio Tests for Regression Parameters, 395
Other Multivariate Test Statistics, 398
Predictions from Multivariate Multiple Regressions, 399
7.8 The Concept of Linear Regression 401
Prediction of Several Variables, 406
Partial Correlation Coefficient, 409
7.9 Comparing the Two Formulations of the Regression Model 410
Mean Corrected Form of the Regression Model, 410
Relating the Formulations, 412
7.10 Multiple Regression Models with Time Dependent Errors 413
Supplement 7A: The Distribution of the Likelihood Ratio for the Multivariate Multiple Regression Model 418
Exercises 420
References 428

8 *PRINCIPAL COMPONENTS* 430

8.1 Introduction 430
8.2 Population Principal Components 430
Principal Components Obtained from Standardized Variables, 436
Principal Components for Covariance Matrices with Special Structures, 439
8.3 Summarizing Sample Variation by Principal Components 441
The Number of Principal Components, 444
Interpretation of the Sample Principal Components, 448
Standardizing the Sample Principal Components, 449
8.4 Graphing the Principal Components 454
8.5 Large Sample Inferences 456
Large Sample Properties of $\hat{\lambda}_i$ and $\hat{\mathbf{e}}_i$, 456
Testing for the Equal Correlation Structure, 457
8.6 Monitoring Quality with Principal Components 459
Checking a Given Set of Measurements for Stability, 459
Controlling Future Values, 463
Supplement 8A: The Geometry of the Sample Principal Component Approximation 466
The p-Dimensional Geometrical Interpretation, 468
The n-Dimensional Geometrical Interpretation, 469
Exercises 470
References 480

9 *FACTOR ANALYSIS AND INFERENCE FOR STRUCTURED COVARIANCE MATRICES* 481

9.1 Introduction 481
9.2 The Orthogonal Factor Model 482
9.3 Methods of Estimation 488
The Principal Component (and Principal Factor) Method, 488
A Modified Approach—the Principal Factor Solution, 494
The Maximum Likelihood Method, 495
A Large Sample Test for the Number of Common Factors, 501
9.4 Factor Rotation 504
Oblique Rotations, 512
9.5 Factor Scores 513
The Weighted Least Squares Method, 514
The Regression Method, 516
9.6 Perspectives and a Strategy for Factor Analysis 519
Supplement 9A: Some Computational Details for Maximum Likelihood Estimation 527
Recommended Computational Scheme, 528
Maximum Likelihood Estimators of $\boldsymbol{\rho} = \mathbf{L}_z\mathbf{L}_z' + \boldsymbol{\psi}_z$ 529
Exercises 530
References 538

10 CANONICAL CORRELATION ANALYSIS 539

10.1 Introduction 539
10.2 Canonical Variates and Canonical Correlations 539
10.3 Interpreting the Population Canonical Variables 545
Identifying the Canonical Variables, 545
Canonical Correlations as Generalizations of Other Correlation Coefficients, 547
The First r Canonical Variables as a Summary of Variability, 548
A Geometrical Interpretation of the Population Canonical Correlation Analysis 549
10.4 The Sample Canonical Variates and Sample Canonical Correlations 550
10.5 Additional Sample Descriptive Measures 558
Matrices of Errors of Approximations, 558
Proportions of Explained Sample Variance, 561
10.6 Large Sample Inferences 563
Exercises 567
References 574

11 DISCRIMINATION AND CLASSIFICATION 575

11.1 Introduction 575
11.2 Separation and Classification for Two Populations 576
11.3 Classification with Two Multivariate Normal Populations 584
Classification of Normal Populations When $\Sigma_1 = \Sigma_2 = \Sigma$, 584
Scaling, 589
Fisher's Approach to Classification with Two Populations, 590
Is Classification a Good Idea?, 592
Classification of Normal Populations When $\Sigma_1 \neq \Sigma_2$, 593
11.4 Evaluating Classification Functions 596
11.5 Classification with Several Populations 606
The Minimum Expected Cost of Misclassification Method, 606
Classification with Normal Populations, 609
11.6 Fisher's Method for Discriminating among Several Populations 621
Using Fisher's Discriminants to Classify Objects, 628
11.7 Logistic Regression and Classification 634
Introduction, 634
The Logit Model, 634
Logistic Regression Analysis, 636
Classification, 638
Logistic Regression with Binomial Responses, 640
11.8 Final Comments 644
Including Qualitative Variables, 644
Classification Trees, 644
Neural Networks, 647
Selection of Variables, 648

Testing for Group Differences, 648
Graphics, 649
Practical Considerations Regarding Multivariate Normality, 649
Exercises 650
References 669

12 CLUSTERING, DISTANCE METHODS, AND ORDINATION 671

12.1 Introduction 671
12.2 Similarity Measures 673
Distances and Similarity Coefficients for Pairs of Items, 673
Similarities and Association Measures for Pairs of Variables, 677
Concluding Comments on Similarity, 678
12.3 Hierarchical Clustering Methods 680
Single Linkage, 682
Complete Linkage, 685
Average Linkage, 690
Ward's Hierarchical Clustering Method, 692
Final Comments—Hierarchical Procedures, 695
12.4 Nonhierarchical Clustering Methods 696
K-means Method, 696
Final Comments—Nonhierarchical Procedures, 701
12.5 Clustering Based on Statistical Models 703
12.6 Multidimensional Scaling 706
The Basic Algorithm, 708
12.7 Correspondence Analysis 716
Algebraic Development of Correspondence Analysis, 718
Inertia, 725
Interpretation in Two Dimensions, 726
Final Comments, 726
12.8 Biplots for Viewing Sampling Units and Variables 726
Constructing Biplots, 727
12.9 Procrustes Analysis: A Method for Comparing Configurations 732
Constructing the Procrustes Measure of Agreement, 733
Supplement 12A: Data Mining 740
Introduction, 740
The Data Mining Process, 741
Model Assessment, 742
Exercises 747
References 755

APPENDIX 757

DATA INDEX 764

SUBJECT INDEX 767

Preface

INTENDED AUDIENCE

This book originally grew out of our lecture notes for an "Applied Multivariate Analysis" course offered jointly by the Statistics Department and the School of Business at the University of Wisconsin–Madison. *Applied Multivariate Statistical Analysis*, Sixth Edition, is concerned with statistical methods for describing and analyzing multivariate data. Data analysis, while interesting with one variable, becomes truly fascinating and challenging when several variables are involved. Researchers in the biological, physical, and social sciences frequently collect measurements on several variables. Modern computer packages readily provide the numerical results to rather complex statistical analyses. We have tried to provide readers with the supporting knowledge necessary for making proper interpretations, selecting appropriate techniques, and understanding their strengths and weaknesses. We hope our discussions will meet the needs of experimental scientists, in a wide variety of subject matter areas, as a readable introduction to the statistical analysis of multivariate observations.

LEVEL

Our aim is to present the concepts and methods of multivariate analysis at a level that is readily understandable by readers who have taken two or more statistics courses. We emphasize the applications of multivariate methods and, consequently, have attempted to make the mathematics as palatable as possible. We avoid the use of calculus. On the other hand, the concepts of a matrix and of matrix manipulations are important. We do not assume the reader is familiar with matrix algebra. Rather, we introduce matrices as they appear naturally in our discussions, and we then show how they simplify the presentation of multivariate models and techniques.

The introductory account of matrix algebra, in Chapter 2, highlights the more important matrix algebra results as they apply to multivariate analysis. The Chapter 2 supplement provides a summary of matrix algebra results for those with little or no previous exposure to the subject. This supplementary material helps make the book self-contained and is used to complete proofs. The proofs may be ignored on the first reading. In this way we hope to make the book accessible to a wide audience.

In our attempt to make the study of multivariate analysis appealing to a large audience of both practitioners and theoreticians, we have had to sacrifice a consistency of level. Some sections are harder than others. In particular, we have summarized a voluminous amount of material on regression in Chapter 7. The resulting presentation is rather succinct and difficult the first time through. We hope instructors will be able to compensate for the unevenness in level by judiciously choosing those sections, and subsections, appropriate for their students and by toning them down if necessary.

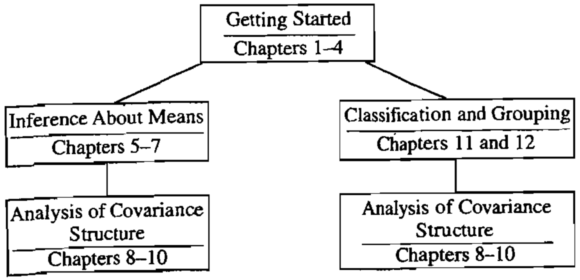
ORGANIZATION AND APPROACH

The methodological "tools" of multivariate analysis are contained in Chapters 5 through 12. These chapters represent the heart of the book, but they cannot be assimilated without much of the material in the introductory Chapters 1 through 4. Even those readers with a good knowledge of matrix algebra or those willing to accept the mathematical results on faith should, at the very least, peruse Chapter 3, "Sample Geometry," and Chapter 4, "Multivariate Normal Distribution."

Our approach in the methodological chapters is to keep the discussion direct and uncluttered. Typically, we start with a formulation of the population models, delineate the corresponding sample results, and liberally illustrate everything with examples. The examples are of two types: those that are simple and whose calculations can be easily done by hand, and those that rely on real-world data and computer software. These will provide an opportunity to (1) duplicate our analyses, (2) carry out the analyses dictated by exercises, or (3) analyze the data using methods other than the ones we have used or suggested.

The division of the methodological chapters (5 through 12) into three units allows instructors some flexibility in tailoring a course to their needs. Possible sequences for a one-semester (two quarter) course are indicated schematically.

Each instructor will undoubtedly omit certain sections from some chapters to cover a broader collection of topics than is indicated by these two choices.



For most students, we would suggest a quick pass through the first four chapters (concentrating primarily on the material in Chapter 1; Sections 2.1, 2.2, 2.3, 2.5, 2.6, and 3.6; and the "assessing normality" material in Chapter 4) followed by a selection of methodological topics. For example, one might discuss the comparison of mean vectors, principal components, factor analysis, discriminant analysis and clustering. The discussions could feature the many "worked out" examples included in these sections of the text. Instructors may rely on diagrams and verbal descriptions to teach the corresponding theoretical developments. If the students have uniformly strong mathematical backgrounds, much of the book can successfully be covered in one term.

We have found individual data-analysis projects useful for integrating material from several of the methods chapters. Here, our rather complete treatments of multivariate analysis of variance (MANOVA), regression analysis, factor analysis, canonical correlation, discriminant analysis, and so forth are helpful, even though they may not be specifically covered in lectures.

CHANGES TO THE SIXTH EDITION

New material. Users of the previous editions will notice several major changes in the sixth edition.

- Twelve new data sets including national track records for men and women, psychological profile scores, car body assembly measurements, cell phone tower breakdowns, pulp and paper properties measurements, Mali family farm data, stock price rates of return, and Concho water snake data.
- Thirty seven new exercises and twenty revised exercises with many of these exercises based on the new data sets.
- Four new data based examples and fifteen revised examples.
- Six new or expanded sections:
    1. Section 6.6 Testing for Equality of Covariance Matrices
    2. Section 11.7 Logistic Regression and Classification
    3. Section 12.5 Clustering Based on Statistical Models
    4. Expanded Section 6.3 to include "An Approximation to the Distribution of T^2 for Normal Populations When Sample Sizes are not Large"
    5. Expanded Sections 7.6 and 7.7 to include Akaike's Information Criterion
    6. Consolidated previous Sections 11.3 and 11.5 on two group discriminant analysis into single Section 11.3

Web Site. To make the methods of multivariate analysis more prominent in the text, we have removed the long proofs of Results 7.2, 7.4, 7.10 and 10.1 and placed them on a web site accessible through *www.prenhall.com/statistics*. Click on "Multivariate Statistics" and then click on our book. In addition, all full data sets saved as ASCII files that are used in the book are available on the web site.

Instructors' Solutions Manual. An Instructors Solutions Manual is available on the author's website accessible through *www.prenhall.com/statistics*. For information on additional for-sale supplements that may be used with the book or additional titles of interest, please visit the Prentice Hall web site at *www.prenhall.com*.