

# Applied Multivariate Statistical Analysis

~.~

00 · · · - - 53- .



# **Applied Multivariate Statistical Analysis**

RICHARD A. JOHNSON University of Wisconsin—Madison

DEAN W. WICHERN Texas A&M University

1000 - 10000 - 1000 - 1000 - 1000 - 1000 - 1000 - 1000 - 1000 - 1000 - 1



Upper Saddle River, New Jersey 07458



#### brary of Congress Cataloging-in-Publication Data

hnson, Richard A. Statistical analysis/Richard A. Johnson.—6<sup>th</sup> ed. Dean W. Winchern p. cm. Includes index. ISBN 0-13-187715-1 1. Statistical Analysis

CIP Data Available

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

PEARSON Prentice Hall © 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 Education 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 **1** ASPECTS OF MULTIVARIATE ANALYSIS Introduction 1 1.11.2 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

- 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

49

XV

1

	Contents
VIII	Contents

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

## 3 SAMPLE GEOMETRY AND RANDOM SAMPLING

111

- Introduction 111 3.1 The Geometry of the Sample 111 3.2 Random Samples and the Expected Values of the Sample Mean and 3.3 Covariance Matrix 119 Generalized Variance 123 3.4 Situations in which the Generalized Sample Variance 1s Zero, 129
- Generalized Variance Determined by |R|and Its Geometrical Interpretation, 134 Another Generalization of Variance, 137 Sample Mean, Covariance, and Correlation 3.5
- As Matrix Operations 137 Sample Values of Linear Combinations of Variables 140 3.6 Exercises 144
  - References 148

#### THE MULTIVARIATE NORMAL DISTRIBUTION 4

149

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

References 208

5	INFE	RENCES ABOUT A MEAN VECTOR	210
	5.1	Introduction 210	
	5.2	The Plausibility of $\mu_0$ as a Value for a Normal Population Mean 210	
	5.3	Hotelling's T <sup>2</sup> 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	<ul> <li>Multivariate Quality Control Charts 239</li> <li>Charts for Monitoring a Sample of Individual Multivariate Observations for Stability, 241</li> <li>Control Regions for Future Individual Observations, 247</li> <li>Control Ellipse for Future Observations, 248</li> <li>T<sup>2</sup>-Chart for Future Observations, 248</li> <li>Control Charts Based on Subsample Means, 249</li> <li>Control Regions for Future Subsample Observations, 251</li> </ul>	
	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 <i>p</i> -Dimensional Ellipsoids 258	
		Exercises 261	
		References 272	
6	сом	IPARISONS 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 $\Sigma_1 \neq \Sigma_2$ , 291	

- An Approximation to the Distribution of  $T^2$  for Normal Populations When Sample Sizes Are Not Large, 294
- **Comparing Several Multivariate Population Means** 6.4 (One-Way Manova) 296 Assumptions about the Structure of the Data for One-Way MANOVA, 296

Contents

## 210

ix

Contents

xi

430

481

Contents

7

6.5 6.6 6.7 6.8 6.9 6.10	A Summary of Univariate ANOVA, 297 Multivariate Analysis of Variance (MANOVA), 301 Simultaneous Confidence Intervals for Treatment Effects 308 Testing for Equality of Covariance Matrices 310 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 Profile Analysis 323 Repeated Measures Designs and Growth Curves 328 Perspectives and a Strategy for Analyzing Multivariate Models 332 Exercises 337	
	References 358	
MUL	TIVARIATE LINEAR REGRESSION MODELS	36
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, 307	
- 4	Inforences A bout the Regression Model 370	
/.4	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 20, 370	
- /	Forecasting a New Observation at 20, 577	
7.6	Does the Model Fit?, 381	
	Leverage and Influence, 384	
	Additional Problems in Linear Regression, 384	
7.7	Multivariate Multiple Regression 387	
	Likelihood Ratio Tesis for Regression Literinetergers	
	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	
	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	

360

#### PRINCIPAL COMPONENTS 8

- Introduction 430 8.1
- 8.2 Population Principal Components 430 Principal Components Obtained from Standardized Variables, 436 Principal Components for Covariance Matrices with Special Structures, 439
- Summarizing Sample Variation by Principal Components 441 8.3 The Number of Principal Components, 444 Interpretation of the Sample Principal Components, 448 Standardizing the Sample Principal Components, 449
- Graphing the Principal Components 454 8.4

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

Monitoring Quality with Principal Components 459 8.6 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

- 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  $\rho = L_z L'_z + \psi_z$  529 Exercises 530 References 538

Contents xii

11

## 10 CANONICAL CORRELATION ANALYSIS

Contents	xiii
----------	------

CANONICAL CORRELATION ANALYSIS     539     Tarting for Champ Difference, 648       101     Introduction 539     Proteinal Considerations Regarding Multivations Normality, 649       102     Canonical Correlations 539     References 669       103     Interpreting the Population Canonical Viriables 55     References 669       104     The Sample Canonical Viriables 545     References 669       105     Actional Viriables 345     12       106     The Sample Canonical Viriables 345     12       107     The Sample Canonical Viriables 345     12       108     The Sample Canonical Viriables 345     12       109     Activities and Sample Canonical Viriables 350     12       1015     Additional Sample Descriptive Measures 53     16       1016     Large Sample Inferences 543     12       1017     The Sample Canonical Viriables 375     12       1018     Additional Sample Descriptive Measures 53     16       1019     Additional Sample Viriance, 504     12       1011     Introduction 575     12       1012     Separation and Markon 400     640       111     Introduction 575     12       112     Separation and Classification with the Populations, 590     12       113     Canonical Viriable 40     640       114     Eval	-		¢' #				;
10.1       Introduction       539       Practical Constantions Regarding Multivariate Normality, 649         10.2       Canonical Controlations       Starting the Population Canonical Correlations       Starting Constantions Regarding Multivariate Normality, 649         10.3       Interpreting the Population Canonical Variables       Starting Constantions       Starting Multivariate Normality, 649         10.3       Interpreting the Population Canonical Variates and Sample       Constantions       Starting Constantions	CANC	ONICAL CORRELATION ANALYSIS	539			Testing for Group Differences, 648 Graphics, 649	
10.2     Canonical Variates and Canonical Variables 353     Exercises 650       10.3     Interpreting the Population Canonical Variables 353     References 660       10.4     Interpreting the Population Canonical Variables 354     References 660       10.7     The First 4 Canonical Variables 354     12.1     Introduction 671     571       10.7     The Start 4 Canonical Variables 374     12.1     Introduction 671     12.2       10.7     The Start 4 Canonical Variables 374     12.1     Introduction 671     12.2       10.8     A Geometania Intelling Variance and Sample     12.2     Similarity of Samilarity of Sa	10.1	Introduction 539	· 年 五 五			Practical Considerations Regarding Multivariate Normality, 649	•
10.3       Interpreting the Population Canonical Variables 545       References 669         10.3       Interpreting the Populations as Generalizations of Other Correlation Configurations as Generalizations of Other Correlation Configurations as Constraints       671         10.4       The Sample Canonical Variables as a Sample Canonical Variables as a Sample Contents       122         10.4       The Sample Canonical Variables as a Sample Canonical Variables as a Sample Contents       121         10.4       The Sample Canonical Variables as Sample Canonical Variables as Sample Contents       122         10.4       The Sample Canonical Variables as Sample Canonical Variables and Sample       122         10.5       Additional Sample Descriptive Measures 538       123         10.6       Large Sample Inferences 501       123         10.6       Large Sample Inferences 504       224         Postering Methods 695       Sample Linkage, 682         Exercises 567       125       Classification of Two Populations 576         11.1       Introduction 675       12.5       Classification of New S <sub>1</sub> = S <sub>1</sub> = S <sub>1</sub> + S <sub>1</sub> 123       Classification with Two Nopulations 576       12.6       12.6       Multimensional Scaling: 706         11.1       Introduction 670       Sample Inferences 732       Large Sample Multimensional Scaling: 732       12.6	10.2	Canonical Variates and Canonical Correlations 539	14 - A			Exercises 650	
of Other Control Configuration Configurations, 547       1         The Featr Connord Variables as Summary of Variability, 548       1         A Geometrical Interpretation of the Population Cardinical Correlations, 740       1         Connord Variables as Summary of Variability, 548       1         A Geometrical Interpretation of the Population Cardinical Correlations 550       1         Canonical Correlations 550       1         Canonical Correlations 550       1         Canonical Statutes and Association Measures Form of Approximations, 558       1         Proportions of Explained Sample Internes, 563       1         Exercises 567       1         DISCOMMATION AND CLASSIFICATION       575         Constitution of Consult Variables, 677       1         Proportions of Explained Sample Internes, 563       1         Exercises 567       1         DISCOMMATION AND CLASSIFICATION       575         T12       Separation and Classification for Two Populations, 584         Cansification with Two Numal Populations 576       12.5         Cassification with Now Num X 1, ± 2, 5.393       12.6         Cassification with Now Num X 1, ± 2, 5.393       12.6         Cassification with Now Num X 1, ± 2, 5.393       12.6         Cassification with Neumal Toppulations 606       12.8	10.3	Interpreting the Population Canonical Variables 545 Identifying the Canonical Variables, 545 Comprised Correlations as Generalizations		13	cus	References 669	671
The First Canonical Variables as & Summary of Variables, 548       12.1       Introduction 611         A Geometrial Interpretation of the Population Canonical       12.2       Similarity Exaces       673         Of The Sample Canonical Variables as & Summary of Variables, 570       12.2       Similarity Exaces       673         Of The Sample Canonical Variables, 570       12.3       Introduction 671       12.2       Similarity Exaces       673         Observation Analysis       580       10.6       Large Sample Interences       563       20.5<		of Other Correlation Coefficients, 547		12	(20)	TENING, DISTANCE METHODS, AND ORDINATION	071
A Geometrical Interpretation of the Population Cannual Correlation Analysis 5-9     12.     Similarity Measures 673       104     The Sample Canonical Variates and Sample Canonical Correlations 550     12.     Similarity Measures 673       105     Additional Sample Descriptive Measures 558     12.     Similarity Measures 073       105     Additional Sample Descriptive Measures 558     12.     Similarity Measures 073       106     Large Sample Inferences 563     12.     Similarity Measures 073       107     Exercisas 567     12.     Nonbiaractinical Clustering Method. 602       112.     Stransitication with Provemations, 575     12.     Nonbiaractinical Clustering Method. 606       113.     Classification with Two Multivatian Normal Populations 576     12.     Nonbiaractinical Clustering Method. 606       114.     Evaluating Classification with Two Populations 475     12.     Nutriting Classification and Classification with Two Populations 476       115.     Classification with Went Ty at \$2, \$3.     584       Classification with Went Ty at \$2, \$3.     12.     Nutriting Classification and Networks, 716       114.     Evaluating Classification With Two Populations 606     12.     Biplots for Viewing Sampling Units and Variables 726       115.     Classification with Networks 606     12.     Biplots for Viewing Sampling Units and Variables 726       116.     Fisher's Method for Discriminati		The First r Canonical Variables as a Summary of Variability, 548			12.1	Introduction 671	
10.4     The Sample Canonical Variates and Sample       Canonical Correlations 550     Similarities and Association Measures       Canonical Correlations 550     For Sint of Approximations, 558       Marices of Forts of Approximations, 558     Concluding Comments on Similarity, 678       Marices of Forts of Approximations, 558     12.3       Marices of Forts of Approximations, 558     12.3       Inter Canonical Clustering Methods, 680     Single Inferences 573       References 574     575       DISCRMINATION AND CLASSIFICATION     575       Classification and Classification for Two Populations 576     12.4       11.1     Introduction 575       11.2     Separation of Normal Populations 584       Classification with Two Multivariate Normal Populations 584       Classification of Normal Populations 584       Classification with Normal Populations 584       Classification with Normal Populati		A Geometrical Interpretation of the Population Canonical Correlation Analysis 549			12.2	Similarity Measures 673 Distances and Similarity Coefficients for Pairs of Items, 673	
10.5       Additional Sample Descriptive Measures 558       Concluding Comments on Simulation, 0:0         Marrices of Errors of Appointmitons, 556       Single Linkage, 682         Complete Inferences 563       Complete Linkage, 683         Exercises 567       References 574         DISCRIMINATION AND CLASSIFICATION       575         DISCRIMINATION	10.4	The Sample Canonical Variates and Sample Canonical Correlations 550				Similarities and Association Measures for Pairs of Variables, 677	
Marices of Errors of Approximation, 558     Proportion of Explained Sample Variance, 501       Proportion of Explained Sample Variance, 501     Single Linkage, 603       Large Sample Inferences     563       Exercises 667     References       References     574       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 576       11.4     Separation and Classification with Two Populations 576       11.5     Classification of Normal Populations When X <sub>1</sub> = X <sub>2</sub> = X.584       Sodim, 559     Sodim, 78       Picher's Approach to Classification with Two Populations 596       11.5     Classification of Normal Populations 696       11.6     Esher's Method for Discriminating Classification Metod, 606       Classification with Normal Populations 606     The Minimum Expected Caref Miceacient Metod, 606       Classification with Normal Populations 606     Constructing Bloots 702       11.4     Evaluations 614     Process 741       11.5     Classification of Namal Populations 606     Constructing Bloots for Viewing Sampling Units and Variables 726       Classification with Normal Populations 606     Constructing Bloots 732       11.6     Esher's Method for Discriminating among Several Populations 624     Construct	10.5	Additional Sample Descriptive Measures 558			10.0	Concluding Comments on Similarity, 0/8	
10.6     Large Sample Inferences 563     Average Linkage, 630       Exercises 567     Average Linkage, 630       References 574     Separation and Classification for Two Populations 576       11.1     Introduction 575       11.2     Separation and Classification of Normal Populations 584       Classification of Normal Populations 596       11.4     Evaluating Classification with Two Populations 606       The Evaluating Classification on Method, 606       Classification with Normal Populations 606       Classification with Normal Populations 606       The Evaluating Classification functions 596       11.5     Classification with Normal Populations 606       11.6     Fisher's Discrimination Classification 634       Introduction, 634     Introduction, 634       Introduction, 64     Introduction, 634       Introduction, 64     <		Matrices of Errors of Approximations, 558 Proportions of Explained Sample Variance, 561			12.3	Single Linkage, 682	
Exercises 567Ward S Hierarchical Clustering Method, 692References 574Final Comments – Hierarchical Procedures, 695DISCRIMINATION AND CLASSIFICATION57511.1Introduction 57511.2Reparation and Classification for Two Populations 57611.3Classification of Norma Populations 57611.4Classification of Norma Populations 57611.5Classification of Norma Populations 57611.6Classification of Norma Populations 584Classification of Norma Populations SellClassification of Norma Populations SellII.4Evaluating Classification Functions SellII.5Classification of Norma Populations 606The Basic Algorithm, 726Classification Normal Populations 606Classification and Classification 634II.6Fiber Discriminants to Classification 644Introduction, 541Interduction, 542Interduction, 543Classification 634Interduction, 644Interduction, 644Inclusting Qualitative Viribles, 644Neural Networks, 647Neural Networks, 647Neural Networks, 648Subject INDEXSubject INDEXSubject INDEX	10.6	Large Sample Inferences 563				Average Linkage, 605	
ReferencesFinal Comments—Hierarchical Procedures, 695DISCRIMINATION AND CLASSIFICATION575Final Comments—Nonhierarchical Clustering Methods11.1Introduction57512.4Nonhierarchical Clustering Methods11.2Separation and Classification for Two Populations57612.5Clustering Based on Statistical Models11.3Classification with Now Multivariate Normal Populations58421.6Multidimensional Scaling70611.3Classification with Two Populations, 59012.6Correspondence Analysis, 716Algebraic Development of Correspondence Analysis, 71811.4Evaluating Classification of Normal Populations when $\mathbf{x}_1 \neq \mathbf{x}_2$ , 59312.7Correspondence Analysis, 71611.4Evaluating Classification with Two Populations, 59012.8Biplots for Viewing Sampling Units and Variables72611.5Classification with Several Populations60612.8Biplots for Viewing Sampling Units and Variables72611.6Fisher's Method for Discriminating among Several Populations, 60012.9Procrustes Analysis: A Method for Constructing Biplos, 727Method Algebraic Development, 72011.4Logistic Regression and Classification of Sol Logistic Regression and Classification 634Appendent Algebraic Development, 720Supplement 12A: Data Mining11.6Fisher's Discriminating Logistic Regression and Classification 634Appendent Algebraic Development, 740The Data Mining 74011.7Logistic Regression and Classification 634Data Mining Process, 741Model Assessment, 742		Exercises 567				Ward's Hierarchical Clustering Method, 692	
DISCRIMMATION AAND CLASSIFICATION57512.4Nonhierarchical Clustering Methods 696 K-means Method, 696 Final Comments—Nonhierarchical Procedures, 7011.1Introduction 57512.5Clustering Based on Statistical Models 7031.2Separation and Classification for Two Populations 57612.5Clustering Based on Statistical Models 7031.3Classification on With Two Multivariate Normal Populations 584 Classification of Normal Populations When $\Sigma_1 = \Sigma_2 = \Sigma_3 84$ Scaling, 389 Ficher's Approach to Classification With Two Populations, 590 Is Classification of Normal Populations When $\Sigma_1 = \Sigma_2 = \Sigma_3 2$ Classification of Normal Populations, 59612.6Multifumensional Scaling 706 The Basic Algorithm, 7081.4Evaluating Classification Functions 59612.7Correspondence Analysis, 716 Algebraic Development of Correspondence Analysis, 718 Interving Classification on Method, 606 Classification with Normal Populations, 59612.8Biplots for Viewing Sampling Units and Variables 726 Constructing Biplots, 7271.5Classification with Several Populations, 60912.9Procrustes Analysis: A Method for Comparing Configurations, 732 Constructing Biplots, 7271.6Fisher's Method for Discriminating among Several Populations, 643 Logistic Regression and Classification 634 Logistic Regression and Classification 634 Logistic Regression Analysis, 646XPPENDIX7571.7Classification, 638 Logistic Regression Analysis, 646APPENDIX7571.8Final Comments 644 Logistic Regression Analysis, 646APPENDIX7571.8Final Comments 644 Logistic Regression on Hibiton, 638APP		References 574				Final Comments—Hierarchical Procedures, 695	
<ul> <li>11.1 Introduction 575</li> <li>12. Separation and Classification for Two Populations 576</li> <li>13. Classification with Two Multivariate Normal Populations 584</li> <li>13. Classification of Normal Populations When Σ<sub>1</sub> = Σ<sub>2</sub> = Σ, 584</li> <li>14. Scaling, 389</li> <li>15. Fisher's Approach to Classification Method, 606</li> <li>15. Classification with Several Populations 606</li> <li>16. Fisher's Method for Discriminanting amount of Misclassification 609</li> <li>17. Logistic Regression and Classification 634</li> <li>17. Logistic Regression and Classification 634</li> <li>17. Logistic Regression and Classification 634</li> <li>18. Final Comments 644</li> <li>Neural Networks 647</li> <li>Neural Networks 648</li> <li>SubJECT INDEX</li> </ul>	DISC	RIMINATION AND CLASSIFICATION	575		12.4	Nonhierarchical Clustering Methods 696 K-means Method, 696	
<ul> <li>Separation and Classification for Two Populations 576</li> <li>Classification with Two Multivariate Normal Populations 584</li> <li>Classification with Two Multivariate Normal Populations 584</li> <li>Classification of Normal Populations When X<sub>1</sub> = X<sub>2</sub> = X, 584</li> <li>Scating, 589</li> <li>Fisher's Approach to Classification with Two Populations 590</li> <li>In Evaluating Classification of Normal Populations 596</li> <li>Classification of Normal Populations 596</li> <li>Classification of Normal Populations 606</li> <li>Classification with Several Populations 606</li> <li>Fisher's Method for Discriminating among Several Populations 621</li> <li>Using Fisher's Discriminants to Classify Objects 628</li> <li>Logistic Regression with Binomial Responses, 640</li> <li>Final Comments 644</li> <li>Inclusing Qualitative G44</li> <li>Classification of Variables, 644</li> <li>Subject Index 644</li> <li>Neural Networks 647</li> <li>Subject Index 648</li> <li>Subject Index 546</li> <li>SUBJECT INDEX</li> </ul>	11.1	Introduction 575			125	Clustering Based on Statistical Models 703	
1.3Classification with Two Multivariate Normal Populations5841.3Multivariate Normal Populations $700$ Classification of Normal Populations When $\Sigma_1 = \Sigma_2 = \Sigma$ , 5841.3Multivariate Algorithm, 708Classification of Normal Populations with Two Populations, 5901.4Multivariate Algorithm, 708Is Classification of Normal Populations When $\Sigma_1 = \Sigma_2$ , 5931.2Correspondence Analysis, 716Algebraic Development of Correspondence Analysis, 718Algebraic Development of Correspondence Analysis, 718Institution Sistification of Normal Populations5961.811.5Classification with Several Populations, 60612.8Classification with Normal Populations, 60612.8Classification with Normal Populations, 60612.8Classification with Normal Populations, 60612.9Procurster Method for Discriminating among Several Populations 621Supplement 12.4: Dat Mining, 740Using Fisher's Discrimination Classify Objects, 628Supplement 12.4: Dat Mining, 74011.7Logistic Regression and Classify Objects, 628Supplement 12.4: Data Mining, 74011.8Final Comments, 634The Data Mining Process, 741Minduition Statis, 636Exercises, 747Classification, 634AppenDIXThe Isola Regression with Binomial Responses, 640AppenDIXNeural Networks, 647SubjECT INDEXSelection of Variables, 648SUBJECT INDEX	11.2	Separation and Classification for Two Populations 576	100 B		12.5	Multidimensional Scaling 706	
Classification of Normal Populations When $\Sigma_1 = \Sigma_2 = \Sigma$ , 58412.7Correspondence Analysis 716 Algebraic Development of Correspondence Analysis, 718 Inertia, 72512.7Correspondence Analysis 716 Algebraic Development of Correspondence Analysis, 718 Inertia, 725 Interpretation in Two Dimensions, 726 Final Comments, 72611.4Evaluating Classification Functions 59612.5Classification of Normal Populations, 606 Classification with Neveral Populations 606 Classification with Normal Populations, 60912.6Fisher's Method for Discriminating among Several Populations 621 Using Fisher's Discriminants to Classification 634 Introduction, 634 Classification of Algebraic Development, 644 Logistic Regression and Classification 638 Logistic Regression with Binomial Responses, 64011.8Final Comments 644 Meural Networks, 647 Selection of Variables, 64812.7Correspondence Analysis 716 Algebraic Development of Correspondence Analysis, 718 Interpretation in Two Dimensions, 726 Final Comments, 72612.7Classification with Several Populations 606 Constructing Biplots, 72713.6Fisher's Method for Discriminating among Several Populations 621 Using Fisher's Discriminants to Classification 634 Introduction, 634 Logistic Regression and Classification 634 Logistic Regression Analysis, 636 Classification 74214.7Logistic Regression Analysis, 636 Classification 74215.8Final Comments 644 Meural Networks, 647 Selection of Variables, 64816.9SUBJECT INDEX17.9SubJect INDEX	11.3	Classification with Two Multivariate Normal Populations 584			12.0	The Basic Algorithm, 708	
Scaling, 389Algebraic Development of Correspondence Analysis, 718 Inertia, 723Fisher's Approach to Classification of Normal Populations, 590 Is Classification of Normal Populations When Σ <sub>1</sub> ≠ Σ <sub>2</sub> , 593Algebraic Development of Correspondence Analysis, 718 Inertia, 72311.4Evaluating Classification of Normal Populations For Classification with Several Populations 606 Classification with Several Populations, 60912.811.6Fisher's Method for Discriminating among Several Populations 621 Using Fisher's Discriminants to Classify Objects, 62812.911.7Logistic Regression and Classification 634 Inroduction, 634 Classification The Logi Model, 634 Logistic Regression and Classification Face 644 Incoduction, 76812.911.8Final Comments, 72675711.8Final Comments, 726 Constructing the Process, 741 Model Assessment, 742 Exercises 75575711.7Logistic Regression and Classification 634 Increduction, 638 Logistic Regression vith Binomial Responses, 64075711.8Final Comments, 644 Inclusing Qualitative Variables, 644 Classification Trees, 644 Neural Networks 647 Selection of Variables, 64850811.8Final Comments, 64850811.9SUBJECT INDEX767		Classification of Normal Populations When $\Sigma_1 = \Sigma_2 = \Sigma_5 584$			12.7	Correspondence Analysis 716	
Inertia, 725       Inertia, 725         Is Classification of Normal Populations When ∑₁ ≠ ∑₂, 593       Inertia, 725         11.4       Evaluating Classification Functions 596       Inertia, 725         11.5       Classification with Several Populations 606       Inertia, 726         The Minimum Expected Cost of Misclassification Method, 606       Classification with Normal Populations, 609       Inertia, 726         11.6       Fisher's Method for Discriminating among Several Populations 621       Constructing Biplots, 727       Method         11.7       Logistic Regression and Classification 634       Supplement 12A: Data Mining 740       Introduction, 740         11.7       Logistic Regression Analysis, 636       Exercises 747       Model, 634         Logistic Regression Vith Binomial Responses, 640       APPENDIX       757         11.8       Final Comments 644       DATA INDEX       764         Neural Networks, 647       SUBJECT INDEX       767		Scaling, 589				Algebraic Development of Correspondence Analysis, 718	
Classification of Normal Populations When X₁ ≠ X₂, 593       Interpretation in two Dimensions, 726         11.4       Evaluating Classification Functions 596       12.8       Biplots for Viewing Sampling Units and Variables 726         11.5       Classification with Normal Populations, 600       12.8       Biplots for Viewing Sampling Units and Variables 726         11.6       Fisher's Method for Discriminating among Several Populations 621       12.9       Procrustes Analysis: A Method for Constructing the Procrustes Measure of Agreement, 733         11.7       Logistic Regression and Classification 634       The Data Mining 740       Introduction, 740         11.7       Logistic Regression Analysis, 636       Exercises 747       References 755         11.8       Final Comments 644       APPENDIX       757         11.8       Final Comments 644       5UBJECT INDEX       764		Is Classification a Good Idea?, 592				Inertia, 725	
11.4       Evaluating Classification Functions 596       12.8       Biplots for Viewing Sampling Units and Variables 726         11.5       Classification with Several Populations 606       12.8       Biplots for Viewing Sampling Units and Variables 726         11.6       Fisher's Method for Discriminating among Several Populations 609       12.9       Procrustes Analysis: A Method for Constructing the Porcurst Method, 606         11.6       Fisher's Method for Discriminating among Several Populations 621       Supplement 12A: Data Mining 740         Using Fisher's Discriminants to Classify Objects, 628       Introduction, 740         11.7       Logistic Regression and Classification 634       The Logit Model, 634         Introduction, 634       The Logit Model, 634       Exercises 747         Logistic Regression Analysis, 636       Classification with Binomial Responses, 640       For Comments 644         Including Qualitative Variables, 644       DATA INDEX       757         Neural Networks, 647       SUBJECT INDEX       764		Classification of Normal Populations When $\Sigma_1 \neq \Sigma_2$ , 593	5			Interpretation in Iwo Dimensions, 720 Final Comments 726	
11.5       Classification with Several Populations 606       Classification With Several Populations 606         The Minimum Expected Cost of Misclassification Method, 606       Classification With Normal Populations, 609         11.6       Fisher's Method for Discriminanting       Supplement 122.         among Several Populations 621       Supplement 122.       Constructing the Procrustes Measure of Agreement, 733         Using Fisher's Discriminants to Classify Objects, 628       Supplement 12A: Data Mining 740       Introduction, 740         11.7       Logistic Regression and Classification 634       The Data Mining Process, 741       Model Assessment, 742         Logistic Regression Analysis, 636       Exercises 747       References 755         Logistic Regression with Binomial Responses, 640       APPENDIX       757         11.8       Final Comments 644       DATA INDEX       764         Neural Networks, 647       SUBJECT INDEX       767	11.4	Evaluating Classification Functions 596	Ē		12.8	Riplots for Viewing Sampling Units and Variables 726	
The Minimum Expected Cost of Misclassification Method, 60% Classification with Normal Populations, 60912.9Procrustes Analysis: A Method for Comparing Configurations, 732 Constructing the Procrustes Measure of Agreement, 733 	11.5	Classification with Several Populations 606	2 		12.0	Constructing Biplots, 727	
Classification with Worman Formating among Several Populations 621 Using Fisher's Discriminants to Classify Objects, 628for Comparing Configurations 732 Constructing the Procrustes Measure of Agreement, 733 		The Minimum Expected Cost of Misclassification Method, 000	alita ya		12.9	Procrustes Analysis: A Method	
11.6       Fisher's Method 101 Discomminants of Agreement, 733         among Several Populations 621       Supplement 12A: Data Mining 740         Using Fisher's Discriminants to Classification 634       Introduction, 740         11.7       Logistic Regression and Classification 634       Model Assessment, 742         Introduction, 634       Exercises 747         Classification, 638       References 755         Logistic Regression with Binomial Responses, 640       APPENDIX       757         11.8       Final Comments 644       DATA INDEX       764         Neural Networks, 647       SUBJECT INDEX       767		Classification with Norman Populations, 005				for Comparing Configurations 732	
antoling Several replacement reprint representation replacement replacement reprint reprin	11.6	Fisher's Method for Discriminating				Constructing the Procrustes Measure of Agreement, 733	
Introduction, 74011.7Logistic Regression and Classification 634The Data Mining Process, 741Introduction, 634Model Assessment, 742The Logit Model, 634Exercises 747Logistic Regression Analysis, 636References 755Classification, 638APPENDIXLogistic Regression with Binomial Responses, 640APPENDIX11.8Final Comments 644DATA INDEXIncluding Qualitative Variables, 644764Selection of Variables, 648SUBJECT INDEX767		Using Fisher's Discriminants to Classify Objects, 628				Supplement 12A: Data Mining 740	
Introduction, 634Model Assessment, 742Introduction, 634Model Assessment, 742Logist Model, 634Exercises 747Logistic Regression Analysis, 636References 755Logistic Regression with Binomial Responses, 640APPENDIX11.8Final Comments 644Including Qualitative Variables, 644DATA INDEXNeural Networks, 647SUBJECT INDEXSelection of Variables, 648767	11.7	Logistic Regression and Classification 634	i t			Introduction, 740 The Data Mining Process 741	
The Logit Model, 634Exercises 747Logistic Regression Analysis, 636References 755Classification, 638References 755Logistic Regression with Binomial Responses, 640APPENDIX11.8Final Comments 644Including Qualitative Variables, 644Including Qualitative Variables, 644DATA INDEXNeural Networks, 647SUBJECT INDEXSelection of Variables, 648767	11.7	Introduction, 634	2			Model Assessment, 742	
Logistic Regression Analysis, 050       References 755         Classification, 638       Logistic Regression with Binomial Responses, 640         11.8       Final Comments 644         Including Qualitative Variables, 644       DATA INDEX         Neural Networks, 647       SUBJECT INDEX         Selection of Variables, 648       767		The Logit Model, 634	1			Exercises 747	
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 <b>DATA INDEX</b> <b>DATA INDEX</b> <b>DATA INDEX</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>757</b> <b>764</b> <b>764</b> <b>764</b> <b>764</b> <b>764</b> <b>764</b> <b>764</b> <b>765</b> <b>767</b> <b>767</b> <b>767</b> <b>767</b> <b>767</b> <b>767</b>		Logistic Regression Analysis, 050 Classification 638		,		References 755	
11.8Final Comments644757Including Qualitative Variables, 644DATA INDEX764Classification Trees, 644Perrol Networks, 647764Neural Networks, 647SUBJECT INDEX767		Logistic Regression with Binomial Responses, 640	- Shurnyeter				
Including Qualitative Variables, 644 Classification Trees, 644 Neural Networks, 647 Selection of Variables, 648 767	11.8	Final Comments 644			APPE	NDIX	757
Classification Trees, 644 Neural Networks, 647 Selection of Variables, 648 SUBJECT INDEX 767		Including Qualitative Variables, 644			DATA	INDEX	764
Selection of Variables, 648 767		Classification Trees, 644	1		2.07		
		Selection of Variables, 648	] 		SUBJ	ECT INDEX	767

11.8	Final Comments 644
	Including Qualitative Variables, 644
	Classification Trees, 644
	Neural Networks, 647
	Selection of Variables, 648

## 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 xvi Preface

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. Fach 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 di-

agrams 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*.









