## Multivariate

## With MATLAB

## Professor Dr. Taha Hussein Ali

## 2020



## Part I

Department of Statistics and Informatics, College of Administration and Economics Salahaddin University, Erbil, Kurdistan Region, Iraq

# Multivariate With <br> <br> MATLAB 

 <br> <br> MATLAB}

Prof. Dr. Taha Hussein Ali Alzubaydi

Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University, Erbil, Kurdistan Region, Iraq, 2020

## Introduction

This book has been prepared for the beginners to help them understand basic to advanced functionality of MATLAB. After completing this chapter 1 (Which included an explanation of the Matlab language) you will find yourself at a moderate level of expertise in using MATLAB from where you can take yourself to next levels.
On other side, I have long been fascinated by the interplay of variables in multivariate data and by the challenge of unraveling the effect of each variable. My continuing objective has been to present the power and utility of multivariate analysis in a highly readable format.

Practitioners and researchers in all applied disciplines often measure several variables on each subject or experimental unit. In some cases, it may be productive to isolate each variable in a system and study it separately. Typically, however, the variables are not only correlated with each other, but each variable is influenced by the other variables as it affects a test statistic or descriptive statistic. Thus, in many instances, the variables are intertwined in such a way that when analyzed individually they yield little information about the system. Using multivariate analysis, the variables can be examined simultaneously in order to access the key features of the process that produced them. The multivariate approach enables us to (1) explore the joint performance of the variables and (2) determine the effect of each variable in the presence of the others.
Multivariate analysis provides both descriptive and inferential procedures-we can search for patterns in the data or test hypotheses about patterns of a priori interest. With multivariate descriptive techniques, we can peer beneath the tangled web of variables on the surface and extract the essence of the system. Multivariate inferential procedures include hypothesis tests that (1) process any number of variables without inflating the Type I error rate and (2) allow for whatever intercorrelations the variables possess. A wide variety of multivariate descriptive and inferential procedures is readily accessible in statistical software packages.

## CONTENTS

| 1 | Chapter one: Introduction to MATLAB | 8 |
| :---: | :--- | :---: |
| 1.1 | Introduction | 9 |
| 1.2 | MATLAB's Power of Computational Mathematics | 9 |
| 1.3 | Features of MATLAB | 10 |
| 1.4 | Desktop Basics | 10 |
| 1.5 | Matrices and Vectors | 12 |
| 1.5 .1 | Assignment and Operators | 14 |
| 1.5 .2 | Extracting a Sub-Matrix | 15 |
| 1.5 .3 | Matrix Functions in Matlab | 16 |
| 1.6 | Pre-Defined Variables | 17 |
| 1.7 | Plotting in Matlab | 22 |
| 1.8 | Logical Subscripting | 22 |
| 1.9 | Multidimensional Arrays | 24 |
| 1.10 | Programming in Matlab | 25 |
| 1.10 .1 | Relational Operators | 25 |
| 1.10 .2 | Logical Operators | 26 |
| 1.10 .3 | Conditional Structures | 27 |
| 1.11 | Matlab Iteration Structures | 28 |
| 1.12 | M-Files | 29 |
| 1.12 .1 | M-Files -Scripts | 30 |
| 1.12 .2 | M-Files -Functions | 31 |
| 1.13 | Debugging in Matlab | 32 |
| 1.14 | Advanced Features to Explore |  |
| 1.15 | Descriptive statistics with the Statistics Toolbox of | 33 |
|  | MATLAB | 36 |
| 1.16 | Simulation of linear models | 36 |
| 1.16 .1 | Simulation of simple linear model | 41 |
| 1.16 .2 | Ordinary Least Squares Regression | 45 |
| 1.16 .3 | Simple linear regression in matrix form | 47 |
| 1.16 .4 | Multiple Linear Regression | 50 |
| 1.16 .5 | Multiple linear regression with the Statistics Toolbox of | 50 |
| 1.17 | MATLAB | 53 |
| 1.17 .1 | Simulation of Stochastic processes | 53 |
| 1.17 .2 | Simulation of Bernoulli process | 55 |
|  | Simulan of Random walk |  |


| 1.17 .3 | Simulation of Poisson process | 55 |
| :---: | :--- | :---: |
| 1.17 .4 | Simulation of Autoregressive process | 56 |
| 1.17 .5 | Simulation of Moving average process | 57 |
| 1.18 | Nonlinear Regression | 58 |
| 1.18 .1 | Nonlinear Transformations | 58 |
| 1.18 .2 | Polynomial fitting | 61 |
|  | PROBLEMS | 64 |
| 2 | Chapter two: Tests on One or Two Mean Vectors | 67 |
| 2.1 | MULTIVARIATE VERSUS UNIVARIATE TESTS | 68 |
| 2.2 | TESTS ON $\mu$ WITH $\Sigma$ KNOWN | 69 |
| 2.2 .1 | Review of Univariate Test for $H_{0}: \mu=\mu_{0}$ with $\sigma$ Known | 69 |
| 2.2 .2 | Multivariate Test for $H_{0}: \mu=\mu_{0}$ with $\Sigma$ Known | 70 |
| 2.3 | TESTS ON $\mu$ WHEN $\Sigma$ IS UNKNOWN | 75 |
| 2.3 .1 | Review of Univariate $t$-Test for $H_{0}: \mu=\mu_{0}$ with $\sigma$ Unknown | 75 |
| 2.3 .2 | Hotelling's $T^{2}$-Test for $H_{0}: \mu=\mu_{0}$ with $\Sigma$ Unknown | 76 |
| 2.4 | COMPARING TWO MEAN VECTORS | 81 |
| 2.4 .1 | Review of Univariate Two-Sample $t$-Test | 81 |
| 2.4 .2 | Multivariate Two-Sample $T^{2}$-Test | 82 |
| 2.4 .3 | Likelihood Ratio Tests | 87 |
| 2.5 | TESTS ON INDIVIDUAL VARIABLES CONDITIONAL | 88 |
|  | ON REJECTION OF $H_{0}$ BY THE $T^{2}$-TEST |  |
| 2.6 | COMPUTATION OF $T^{2}$ | 94 |
| 2.6 .1 | Obtaining $T^{2}$ from a MANOVA Program | 94 |
| 2.6 .2 | Obtaining $T^{2}$ from Multiple Regression | 95 |
| 2.7 | PAIRED OBSERVATIONS TEST | 97 |
| 2.7 .1 | Univariate Case | 97 |
| 2.7 .2 | Multivariate Case | 99 |
| 2.8 | TEST FOR ADDITIONAL INFORMATION | 102 |
| 2.9 | PROFILE ANALYSIS | 107 |
| 2.9 .1 | One-Sample Profile Analysis | 108 |
| 2.9 .2 | Two-Sample Profile Analysis | 111 |
|  | PROBLEMS | 120 |
| 3 | Chapter three: Multivariate Analysis of Variance | 129 |
| 3.1 | ONE-WAY MODELS | 130 |
| 3.1 .1 | Univariate One-Way Analysis of Variance $($ ANOVA) | 130 |
|  |  | 4 |


| 3.1.2 | Multivariate One-Way Analysis of Variance Model <br> (MANOVA) | 132 |
| :---: | :---: | :---: |
| 3.1.3 | Wilks’ Test Statistic | 136 |
| 3.1.4 | Roy's Test | 139 |
| 3.1.5 | Pillai and Lawley-Hotelling Tests | 141 |
| 3.1.6 | Unbalanced One-Way MANOVA | 143 |
| 3.1.7 | Summary of the Four Tests and Relationship to $T^{2}$ | 144 |
| 3.1.8 | Measures of Multivariate Association | 151 |
| 3.2 | COMPARISON OF THE FOUR MANOVA TEST STATISTICS | 156 |
| 3.3 | CONTRASTS | 159 |
| 3.3.1 | Univariate Contrasts | 159 |
| 3.3.2 | Multivariate Contrasts | 161 |
| 3.4 | TESTS ON INDIVIDUAL VARIABLES FOLLOWING REJECTION OF $H_{0}$ BY THE OVERALL MANOVA TEST | 167 |
| 3.5 | TWO-WAY CLASSIFICATION | 172 |
| 3.5.1 | Review of Univariate Two-Way ANOVA | 172 |
| 3.5.2 | Multivariate Two-Way MANOVA | 175 |
| 3.6 | Other Models | 186 |
| 3.6.1 | Higher Order Fixed Effects | 186 |
| 3.6 .2 | Mixed Models | 186 |
| 3.7 | Checking on the Assumptions | 189 |
| 3.8 | Profile Analysis | 190 |
| 3.9 | Repeated Measures Designs | 198 |
| 3.9.1 | Multivariate vs. Univariate Approach | 198 |
| 3.9.2 | One-Sample Repeated Measures Model | 202 |
| 3.9.3 | $k$-Sample Repeated Measures Model | 207 |
| 3.9.4 | Computation of Repeated Measures Tests | 209 |
| 3.9 .5 | Repeated Measures with Two Within-Subjects Factors and One Between-Subjects Factor | 210 |
| 3.9.6 | Repeated Measures with Two Within-Subjects Factors and Two Between-Subjects Factors | 219 |
| 3.9.7 | Additional Topics | 221 |
| 3.10 | Growth Curves | 221 |
| 3.10.1 | Growth Curve for One Sample | 221 |


| 3.10 .2 | Growth Curves for Several Samples | 231 |
| :---: | :--- | :---: |
| 3.10 .3 | Additional Topics | 234 |
| 3.11 | Tests on a Sub-vector | 234 |
| 3.11 .1 | Test for Additional Information | 234 |
| 3.11 .2 | Stepwise Selection of Variables | 239 |
|  | PROBLEMS | 240 |
| 4 | Chapter four: Tests on Covariance Matrices | 256 |
| 4.1 | INTRODUCTION | 257 |
| 4.2 | TESTING A SPECIFIED PATTERN FOR $\Sigma$ | 257 |
| 4.2 .1 | Testing $H_{0}: \Sigma=\Sigma_{0}$ | 257 |
| 4.2 .2 | Testing Sphericity | 259 |
| 4.2 .3 | Testing $H_{0}: \Sigma=\sigma^{2}[(1-\rho)$ I+ $\rho$ J $]$ | 263 |
| 4.3 | Tests Comparing Covariance Matrices | 266 |
| 4.3 .1 | Univariate Tests of Equality of Variances | 267 |
| 4.3 .2 | Multivariate Tests of Equality of Covariance Matrices | 268 |
| 4.4 | Tests of Independence | 273 |
| 4.4 .1 | Independence of Two Sub-vectors | 273 |
| 4.4 .2 | Independence of Several Sub-vectors | 278 |
| 4.4 .3 | Test for Independence of All Variables | 284 |
|  | PROBLEMS | 286 |
| 5 | Chapter five: Discriminant Analysis: <br>  <br> Description of Group Separation | 292 |
| 5.1 | INTRODUCTION | 293 |
| 5.2 | The Discriminant function for two Groups | 294 |
| 5.3 | RELATIONSHIP BETWEEN TWO-GROUP DISCRIMIN- | 300 |
| ANT ANALYSIS AND MULTIPLE REGRESSION |  |  |
| 5.4 | DISCRIMINANT ANALYSIS FOR SEVERAL GROUPS | 302 |
| 5.4 .1 | Discriminant Functions | 302 |
| 5.4 .2 | A Measure of Association for Discriminant Functions | 311 |
| 5.5 | STANDARDIZED DISCRIMINANT FUNCTIONS | 314 |
| 5.6 | TESTS OF SIGNIFICANCE | 318 |
| 5.6 .1 | Tests for the Two-Group Case | 318 |
| 5.7 | INTERPRETATION OF DISCRIMINANT FUNCTIONS | 324 |
| 5.7 .1 | Standardized Coefficients | 325 |
| 5.7 .2 | Partial $F$-Values | 326 |
|  |  |  |


| 5.7 .3 | Correlations between Variables and Discriminant Functions | 327 |
| :---: | :--- | :---: |
| 5.7 .4 | Rotation | 328 |
| 5.8 | SCATTER PLOTS | 328 |
| 5.9 | STEPWISE SELECTION OF VARIABLES | 332 |
|  | PROBLEMS | 343 |
| 6 | Chapter six: Classification Analysis: <br> Allocation of Observations to Groups | 346 |
| 6.1 | INTRODUCTION | 347 |
| 6.2 | CLASSIFICATION INTO TWO GROUPS | 348 |
| 6.3 | CLASSIFICATION INTO SEVERAL GROUPS | 354 |
| 6.3 .1 | Equal Population Covariance Matrices: <br> Linear Classification Functions | 354 |
| 6.3 .2 | Unequal Population Covariance Matrices: <br> Quadratic Classification Functions | 359 |
| 6.4 | ESTIMATING MISCLASSIFICATION RATES | 360 |
| 6.5 | IMPROVED ESTIMATES OF ERROR RATES | 368 |
| 6.5 .1 | Partitioning the Sample | 368 |
| 6.5 .2 | Holdout Method | 368 |
| 6.6 | SUBSET SELECTION | 372 |
| 6.7 | NONPARAMETRIC PROCEDURES | 383 |
| 6.7 .1 | Multinomial Data | 383 |
| 6.7 .2 | Classification Based on Density Estimators | 386 |
| 6.7 .3 | Nearest Neighbor Classification Rule | 394 |
|  | PROBLEMS | 399 |
|  | References | 402 |

## Chapter one

## Introduction to <br> 

$$
S_{2}=\left(\begin{array}{llll}
9.1361 & 7.5494 & 4.8639 & 4.1512 \\
7.5494 & 18.604 & 10.225 & 5.4456 \\
4.8639 & 10.225 & 30.039 & 13.494 \\
4.1512 & 5.4456 & 13.494 & 27.996
\end{array}\right)
$$

Table 2.3: Four Psychological Test Scores on 32 Males and 32 Females

| Males |  |  |  | Females |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $y_{1}$ | $y_{2}$ | $y_{3}$ | $y_{4}$ | $y_{1}$ | $y_{2}$ | $y_{3}$ | $y_{4}$ |
| 15 | 17 | 24 | 14 | 13 | 14 | 12 | 21 |
| 17 | 15 | 32 | 26 | 14 | 12 | 14 | 26 |
| 15 | 14 | 29 | 23 | 12 | 19 | 21 | 21 |
| 13 | 12 | 10 | 16 | 12 | 13 | 10 | 16 |
| 20 | 17 | 26 | 28 | 11 | 20 | 16 | 16 |
| 15 | 21 | 26 | 21 | 12 | 9 | 14 | 18 |
| 15 | 13 | 26 | 22 | 10 | 13 | 18 | 24 |
| 13 | 5 | 22 | 22 | 10 | 8 | 13 | 23 |
| 14 | 7 | 30 | 17 | 12 | 20 | 19 | 23 |
| 17 | 15 | 30 | 27 | 11 | 10 | 11 | 27 |
| 17 | 17 | 26 | 20 | 12 | 18 | 25 | 25 |
| 17 | 20 | 28 | 24 | 14 | 18 | 13 | 26 |
| 15 | 15 | 29 | 24 | 14 | 10 | 25 | 28 |
| 18 | 19 | 32 | 28 | 13 | 16 | 8 | 14 |
| 18 | 18 | 31 | 27 | 14 | 8 | 13 | 25 |
| 15 | 14 | 26 | 21 | 13 | 16 | 23 | 28 |
| 18 | 17 | 33 | 26 | 16 | 21 | 26 | 26 |
| 10 | 14 | 19 | 17 | 14 | 17 | 14 | 14 |
| 18 | 21 | 30 | 29 | 16 | 16 | 15 | 23 |
| 18 | 21 | 34 | 26 | 13 | 16 | 23 | 24 |
| 13 | 17 | 30 | 24 | 2 | 6 | 16 | 21 |
| 16 | 16 | 16 | 16 | 14 | 16 | 22 | 26 |
| 11 | 15 | 25 | 23 | 17 | 17 | 22 | 28 |
| 16 | 13 | 26 | 16 | 16 | 13 | 16 | 14 |
| 16 | 13 | 23 | 21 | 15 | 14 | 20 | 26 |
| 18 | 18 | 34 | 24 | 12 | 10 | 12 | 9 |
| 16 | 15 | 28 | 27 | 14 | 17 | 24 | 23 |
| 15 | 16 | 29 | 24 | 13 | 15 | 18 | 20 |
| 18 | 19 | 32 | 23 | 11 | 16 | 18 | 28 |
| 18 | 16 | 33 | 23 | 7 | 7 | 19 | 18 |
| 17 | 20 | 21 | 21 | 12 | 15 | 7 | 28 |
| 19 | 19 | 30 | 28 | 6 | 5 | 6 | 13 |

The sample covariance matrices do not appear to indicate a disparity in the population covariance matrices. (A significance test to check this assumption is carried out in Example 4.3.2, and the hypothesis $H_{0}: \Sigma_{1}=\Sigma_{2}$ is not rejected.) The pooled covariance matrix is

$$
\begin{aligned}
\mathrm{S}_{\mathrm{pl}} & =\frac{1}{32+32-2}\left[(32-1) \mathrm{S}_{1}+(32-1) \mathrm{S}_{2}\right] \\
& =\left(\begin{array}{llll}
7.164 & 6.047 & 5.693 & 4.701 \\
6.047 & 15.89 & 8.492 & 5.856 \\
5.693 & 8.492 & 29.36 & 13.98 \\
4.701 & 5.856 & 13.98 & 22.32
\end{array}\right)
\end{aligned}
$$

By (2.9), we obtain

$$
T^{2}=\frac{n_{1} n_{2}}{n_{1}+n_{2}}\left(\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}\right)^{\prime} \mathrm{S}_{\mathrm{pl}}^{-1}\left(\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}\right)=97.601
$$

From interpolation in Table A.7, we obtain $T_{0.01,4,62}^{2}=15.363$, and we therefore reject $H_{0}: \mu_{1}=\mu_{2}$. See Example 2.5 for a discussion of which variables contribute most to separation of the two groups.

## Example 2.4.2[The program name mt4.m] $\quad$ Applications using MATLAB

clc
clear all
data=[15 172414131412 21;17 153226141214 26; 15142923121921
21; 13121016121310 16;20 172628112016 16;15 21262112914 18;
$1513262210131824 ; 13522221081323 ; 147301712201923 ; 17153027$
111011 27;1717262012182525;17202824141813 26;1515 29241410
25 28; 1819322813168 14;18 18312714813 25;15 142621131623 28;
18173326162126 26;10 141917141714 14;18 213029161615 23;
18213426131623 24; 131730242616 21;16 161616141622 26;
11152523171722 28;16 132616161316 14;16 13232115142026 ;
$181834241210129 ; 1615282714172423 ; 1516292413151820$;
18193223111618 28;18 1633237719 18;17 2021211215728 ;
19193028656 13];y11=data(:,1);y12=data(:,2);y13=data(:,3);
y14=data(:,4);y21=data(:,5);y22=data(:,6);y23=data(:,7);
y24=data(:,8);y24=data(:,8);m1=mean(data(:,1:4))',m2=mean(data(:,5:8))'
$\mathrm{S} 1=\operatorname{cov}($ data(:, $1: 4))^{\prime}, \mathrm{S} 2=\operatorname{cov}($ data(:,5:8))',n1=length(y11);n2=length(y21);
$\mathrm{Spl}=(1 /(\mathrm{n} 1+\mathrm{n} 2-2))^{*}((\mathrm{n} 1-1) * \mathrm{~S} 1+(\mathrm{n} 2-1) * \mathrm{~S} 2)$
$\mathrm{Ts}=((\mathrm{n} 1 * \mathrm{n} 2) /(\mathrm{n} 1+\mathrm{n} 2)) *(\mathrm{~m} 1-\mathrm{m} 2){ }^{\prime *} \mathrm{inv}(\mathrm{Spl}) *(\mathrm{~m} 1-\mathrm{m} 2)$
alfa $=0.01 ; \mathrm{p}=4 ; \mathrm{v}=\mathrm{n} 1+\mathrm{n} 2-\mathrm{p}-1$;
$\operatorname{tabTs}=(f i n v(1-\mathrm{alfa}, \mathrm{p}, \mathrm{v}) *(\mathrm{n} 1+\mathrm{n} 2-2) * \mathrm{p}) / \mathrm{v}$

Ans.

```
m1 =
            15.969
            15.906
            27.188
            22.75
m2 =
            12.344
            13.906
    16.656
    21.938
s1 =
            5.1925
        4.5454
        6.5222
        4.5454
        6.5222
        5.25
        13.184
        6.7601
        6.7601
        6.2661
        6.2661 14.468 16.645
s2 =
            9.1361
            7.5494
            7.5494
                18.604
                    7.5494
                                    10.225
                5.4456
                                    30.039 13.494
                                    4.1512
                                    5.4456
                                    13.494
                                    27.996
Spl =
            7.1643
                            6.0474
                            15.894
                            8.4924
                        5.693 4.7006
            7.1643
            7.1643
                                    8.4924
                                    5.8558
                            5.8558
                                    29.356
                                    13.981
            7.1643
Ts =
tabTs =}\begin{array}{r}{97.601}\\{15.363}
\begin{tabular}{lcr}
4.5454 & 6.5222 & 5.25 \\
13.184 & 6.7601 & 6.2661 \\
6.7601 & 28.673 & 14.468 \\
6.2661 & 14.468 & 16.645 \\
7.5494 & 4.8639 & 4.1512 \\
18.604 & 10.225 & 5.4456 \\
10.225 & 30.039 & 13.494 \\
5.4456 & 13.494 & 27.996 \\
6.0474 & 5.693 & 4.7006 \\
15.894 & 8.4924 & 5.8558 \\
8.4924 & 29.356 & 13.981 \\
5.8558 & 13.981 & 22.321
\end{tabular}
                                    13.981
                                    22.321
```


### 2.4.3: Likelihood Ratio Tests

The maximum likelihood approach to estimation was introduced ; the likelihood function is the joint density of $y_{1}, y_{2}, \cdots, y_{n}$. The values of the parameters that maximize the likelihood function are the maximum likelihood estimators.

The likelihood ratio method of test construction uses the ratio of the maximum value of the likelihood function assuming $H_{0}$ is true to the maximum under $H_{1}$, which is essentially unrestricted. Likelihood ratio tests usually have good power and sometimes have optimum power over a wide class of alternatives.

When applied to multivariate normal samples and $H_{0}: \mu_{1}=\mu_{2}$, the likelihood ratio approach leads directly to Hotelling's $T^{2}$-test in (2.9). Similarly, in the one sample case, the $T^{2}$-statistic in (2.5) is the likelihood ratio test. Thus the $T^{2}$-test, which we introduced rather informally, is the best test according to certain criteria.

## 2.5: TESTS ON INDIVIDUAL VARIABLES CONDITIONAL ON REJECTION OF $H_{0}$ BY THE $T^{2}$-TEST

If the hypothesis $H_{0}: \mu_{1}=\mu_{2}$ is rejected, the implication is that $\mu_{1 j} \neq \mu_{2 j}$ for at least one $j=1,2, \cdots, p$. But there is no guarantee that $H_{0}: \mu_{1 j}=\mu_{2 j}$ will be rejected for some $j$ by a univariate test. However, if we consider a linear combination of the variables, $z=a^{\prime} y$, then there is at least one coefficient vector a for which

$$
\begin{equation*}
t(\mathrm{a})=\frac{\bar{z}_{1}-\bar{z}_{2}}{\sqrt{\left(1 / n_{1}+1 / n_{2}\right) s_{z}^{2}}} \tag{2.12}
\end{equation*}
$$

will reject the corresponding hypothesis $H_{0}: \mu_{z_{1}}=\mu_{z_{2}}$ or $H_{0}: \mathrm{a}^{\prime} \mu_{1}=\mathrm{a}^{\prime} \mu_{2}$. $\bar{z}_{1}=\mathrm{a}^{\prime} \overline{\mathrm{y}}_{1}$ and $\bar{z}_{2}=\mathrm{a} \overline{\mathrm{y}}_{2}$, the variance estimator $s_{z}^{2}$ is the pooled estimator $\mathrm{a}^{\prime} \mathrm{S}_{\mathrm{pl}} \mathrm{a}$. Thus (2.12) can be written as

$$
\begin{equation*}
t(\mathrm{a})=\frac{\mathrm{a}^{\prime} \overline{\mathrm{y}}_{1}-\mathrm{a}^{\prime} \overline{\mathrm{y}}_{2}}{\sqrt{\left[\left(n_{1}+n_{2}\right) / n_{1} n_{2}\right] \mathrm{a}^{\prime} \mathrm{S}_{\mathrm{pl}} \mathrm{a}}} \tag{2.13}
\end{equation*}
$$

Since $t(\mathrm{a})$ can be negative, we work with $t^{2}(\mathrm{a})$. The linear function $z=a^{\prime} y$ is a projection of $\mathbf{y}$ onto a line through the origin. We seek the line (direction) on which the difference $\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}$ is maximized when projected. The projected difference $\mathrm{a}^{\prime}\left(\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}\right)$ [standardized by $\mathrm{a}^{\prime} \mathrm{S}_{\mathrm{pl}} \mathrm{a}$ as in (2.13)] will be less in any other direction than that parallel to the line joining $\bar{y}_{1}$ and $\bar{y}_{2}$. The value of a that projects onto this line, or, equivalently, maximizes $t^{2}(\mathrm{a})$ in (2.13), is (any multiple of)

$$
\begin{equation*}
\mathrm{a}=\mathrm{S}_{\mathrm{pl}}^{-1}\left(\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}\right) \tag{2.14}
\end{equation*}
$$

Since a in (2.14) projects $\bar{y}_{1}-\bar{y}_{2}$ onto a line parallel to the line joining $\overline{\mathrm{y}}_{1}$ and $\overline{\mathrm{y}}_{2}$, we would expect that $t^{2}(\mathrm{a})=T^{2}$, and this is indeed the case (see Problem 2.3).
When $\mathrm{a}=\mathrm{S}_{\mathrm{pl}}^{-1}\left(\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}\right)$, then $z=\mathrm{a}^{\prime} \mathrm{y}$ is called the discriminant function. Sometimes the vector a itself in (2.14) is loosely referred to as the discriminant function.

If $H_{0}: \mu_{1}=\mu_{2}$ is rejected by $T^{2}$ in (2.9), the discriminant function a'y will lead to rejection of $H_{0}: a^{\prime} \mu_{1}=a^{\prime} \mu_{2}$ using (2.13), with $a=S_{p l}^{-1}\left(\bar{y}_{1}-\bar{y}_{2}\right)$. We can then examine each $a_{j}$ in $\mathbf{a}$ for an indication of the contribution of the corresponding $y_{j}$ to rejection of $H_{0}$. This follow-up examination of each $a_{j}$ should be done only if $H_{0}: \mu_{1}=\mu_{2}$ is rejected by $T^{2}$. The discriminant function will appear again in Section 2.6.2 and in Chapters 5 and 6.
We list these and other procedures that could be used to check each variable following rejection of $H_{0}$ by a two-sample $T^{2}$-test:

1. Univariate $t$-tests, one for each variable,

$$
\begin{equation*}
t_{j}=\frac{\bar{y}_{1 j}-\bar{y}_{2 j}}{\sqrt{\left[\left(n_{1}+n_{2}\right) / n_{1} n_{2}\right] s_{j j}}}, \quad j=1,2, \cdots, p \tag{2.15}
\end{equation*}
$$

where $s_{j j}$ is the $j$ th diagonal element of $\mathrm{S}_{\mathrm{p}}$. Reject $H_{0}: \mu_{1 j}=\mu_{2 j}$ if $\left|t_{j}\right|>t_{\alpha / 2, n_{1}+n_{2}-2}$. For confidence intervals on $\mu_{1 j}-\mu_{2 j}$, see Rencher (1998, Section 3.6).
2. To adjust the $\alpha$-level resulting from performing the $p$ tests in (2.15), we could use a Bonferroni critical value $t_{\alpha / 2 p, n_{1}+n_{2}-2}$ for (2.15) (Bonferroni 1936). A critical value $t_{\alpha / 2 p}$ is much greater than the corresponding $t_{\alpha / 2}$, and the resulting overall $\alpha$-level is conservative. Bonferroni critical values $t_{\alpha / 2 p, v}$ are given in Table A.8, from Bailey (1977).
3. Another critical value that could be used with (2.15) is $T_{\alpha, p, n_{1}+n_{2}-2}$, where $T_{\alpha}$ is the square root of $T_{\alpha}^{2}$ from Table A.7; that is, $T_{\alpha, p, n_{1}+n_{2}-2}=\sqrt{T_{\alpha, p, n_{1}+n_{2}-2}^{2}}$. This allows for all $p$ variables to be tested as well as all possible linear combinations, as in (2.13), even linear combinations chosen after seeing the data. Consequently, the use of $T_{\alpha}$ is even more conservative than using $t_{\alpha / 2 p}$; that is, $T_{\alpha, p, n_{1}+n_{2}-2}>t_{\alpha / 2 p, n_{1}+n_{2}-2}$.
4. Partial $F$ - or $t$-tests [test of each variable adjusted for the other variables; see (2.32) in Section 2.8].
5. Standardized discriminant function coefficients (see Section 5.5)
6. Correlations between the variables and the discriminant function (see Section 5.7.3).
7. Stepwise discriminant analysis (see Section 5.9).

The first three methods are univariate approaches that do not use covariances or correlations among the variables in the computation of the test statistic. The last four methods are multivariate in the sense that the correlation structure is explicitly taken into account in the computation.
Method 6, involving the correlation between each variable and the discriminant function, is recommended in many texts and software packages. However, Rencher (1988) has shown that these correlations are proportional to individual $t$ - or $F$-tests (see Section 5.7.3). Thus this method is equivalent to method 1 and is a univariate rather than a multivariate approach. Method 7 is often used to identify a subset of important variables or even to rank the variables according to order of entry. But Rencher and Larson (1980) have shown that stepwise methods have a high risk of selecting spurious variables, unless the sample size is very large.

We now consider the univariate procedures 1,2 , and 3 . The probability of rejecting one or more of the $p$ univariate tests when $H_{0}$ is true is called the overall $\alpha$ or experimentwise error rate. If we do univariate tests only, with no $T^{2}$-test, then the tests based on $t_{\alpha / 2 p}$ and $T_{\alpha}$ in procedures 2 and 3 are conservative (overall $\alpha$ too low), and tests based on $t_{\alpha / 2}$ in procedure 1 are liberal (overall $\alpha$ too high). However, when these tests are carried out only after rejection by the $T^{2}$-test (such tests are sometimes called protected tests), the experimentwise error rates change. Obviously the tests will reject less often (under $H_{0}$ ) if they are carried out only if $T^{2}$ rejects. Thus the tests using $t_{\alpha / 2 p}$ and $T_{\alpha}$ become even more conservative, and the test using $t_{\alpha / 2}$ becomes more acceptable.

Hummel and Sligo (1971) studied the experimentwise error rate for univariate $t$-tests following rejection of $H_{0}$ by the $T^{2}$-test (protected tests). Using $\alpha=0.05$, they found that using $t_{\alpha / 2}$ for a critical value yields an overall $\alpha$ acceptably close to the nominal 0.05 . In fact, it is slightly conservative, making this the preferred univariate test (within the limits of their study). They also compared this procedure with that of performing univariate tests without a prior $T^{2}$-test (unprotected tests). For this case, the overall $\alpha$ is too high, as expected. Table 2.2 gives an excerpt of Hummel and Sligo's results. The sample size is for each of the two samples; the $r^{2}$ in common is for every pair of variables.

Hummel and Sligo therefore recommended performing the multivariate $T^{2}$-test followed by univariate $t$-tests. This procedure appears to have the desired overall $\alpha$ level and will clearly have better power than tests using $T_{\alpha}$ or $t_{\alpha / 2 p}$ as a critical value. Table 2.2 also highlights the importance of using univariate $t$-tests only if the multivariate $T^{2}$-test is significant. The inflated $\alpha$ 's resulting if $t$-tests are used without regard to the outcome of the $T^{2}$-test are clearly evident. Thus among the three univariate procedures (procedures 1, 2, and 3), the first appears to be preferred.
Among the multivariate approaches (procedures 4, 5, and 7), we prefer the fifth procedure, which compares the (absolute value of) coefficients in the discriminant function to find the effect of each variable in separating the two groups of observations. These coefficients will often tell a different story from the univariate tests, because the univariate tests do not take into account the correlations among the variables or the effect of each variable on $T^{2}$ in the presence of the other variables. A variable will typically have a different effect in the presence of other variables than it has by itself. In the discriminant function $z=a^{\prime} y=$ $a_{1} y_{1}+a_{2} y_{2}+\cdots+a_{p} y_{p}$, where $\mathrm{a}=\mathrm{S}_{\mathrm{pl}}^{-1}\left(\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}\right)$, the coefficients $a_{1}, a_{2}, \cdots, a_{p}$ indicate the relative importance of the variables in a multivariate context, something the univariate $t$-tests cannot do. If the variables are not commensurate (similar in scale and variance), the coefficients should be standardized, as in Section 5.5; this allows for more valid comparisons among the variables. Rencher and Scott (1990) provided a decomposition of the information in the standardized discriminant
function coefficients. For a detailed analysis of the effect of each variable in the presence of the other variables, see Rencher (1993; 1998, Sections 3.3.5 and 3.5.3).

Table 2.4: Comparison of Experimentwise Error Rates (Nominal $\alpha=.05$ )

| Sample <br> Size | Number of Variables | Common $r^{2}$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | . 10 | . 30 | . 50 | . 70 |
| Univariate Tests Only ${ }^{\text {a }}$ |  |  |  |  |  |
| 10 | 3 | . 145 | . 112 | . 114 | . 077 |
| 10 | 6 | . 267 | . 190 | . 178 | . 111 |
| 10 | 9 | . 348 | . 247 | . 209 | . 129 |
| 30 | 3 | . 115 | . 119 | . 117 | . 085 |
| 30 | 6 | . 225 | . 200 | . 176 | . 115 |
| 30 | 9 | . 296 | . 263 | . 223 | . 140 |
| 50 | 3 | . 138 | . 124 | . 102 | . 083 |
| 50 | 6 | . 230 | . 190 | . 160 | . 115 |
| 50 | 9 | . 324 | . 258 | . 208 | . 146 |
| Multivariate Test Followed by Univariate Tests ${ }^{\text {b }}$ |  |  |  |  |  |
| 10 | 3 | . 044 | . 029 | . 035 | . 022 |
| 10 | 6 | . 046 | . 029 | . 030 | . 017 |
| 10 | 9 | . 050 | . 026 | . 025 | . 018 |
| 30 | 3 | . 037 | . 044 | . 029 | . 025 |
| 30 | 6 | . 037 | . 037 | . 032 | . 021 |
| 30 | 9 | . 042 | . 042 | . 030 | . 021 |
| 50 | 3 | . 038 | . 041 | . 033 | . 028 |
| 50 | 6 | . 037 | . 039 | . 028 | . 027 |
| 50 | 9 | . 036 | . 038 | . 026 | . 020 |

${ }^{a}$ Ignoring multivariate tests.
${ }^{b}$ Carried out only if multivariate test rejects.
Example 2.5: For the psychological data in Table 2.3, we obtained $\overline{\mathrm{y}}_{1}$, $\bar{y}_{2}$, and $\mathrm{S}_{\mathrm{pl}}$ in Example 2.4.2. The discriminant function coefficient vector is obtained from (2.14) as

$$
\mathrm{a}=\mathrm{S}_{\mathrm{pl}}^{-1}\left(\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}\right)=\left(\begin{array}{c}
0.5104 \\
-0.2033 \\
0.4660 \\
-0.3097
\end{array}\right)
$$

Thus the linear combination that best separates the two groups is

$$
a^{\prime} y=0.5104 y_{1}-0.2033 y_{2}+0.4660 y_{3}-0.3097 y_{4}
$$

in which $y_{1}$ and $y_{3}$ appear to contribute most to separation of the two groups. (After standardization, the relative contribution of the variables changes somewhat; see the answer to Problem 5.7 in Appendix B.)

## Example 2.5[The program name mt5.m] $\quad$ Applications using MATLAB

clc
clear all
data=[15 $17241413141221 ; 1715322614121426 ; 15142923$ $12192121 ; 13121016121310$ 16;20 172628112016 16;15 2126 2112914 18; 15132622101318 24;135 222210813 23;14 730
17122019 23;17 $15302711101127 ; 17172620121825$ 25;17 $20282414181326 ; 1515292414102528 ; 1819322813168$ 14; $181831271481325 ; 15142621131623$ 28;18 1733261621 26 26;10 141917141714 14; 18 $21302916161523 ; 18213426$ $13162324 ; 13173024261621 ; 1616161614162226 ; 111525$ 23171722 28;16 132616161316 14;16 132321151420 26;18 18 $34241210129 ; 16152827141724$ 23;15 162924131518 20;
$1819322311161828 ; 18163323771918 ; 172021211215728$;
19193028656 13];y11=data(:,1);y12=data(:,2);y13=data $(:, 3)$;
y14=data(:,4);y21=data(:,5);y22=data(:,6);y23=data(:,7);
y24=data(:,8);y24=data(:,8);
m1=mean(data(:,1:4))';
m2=mean(data(:,5:8))';
S1=cov(data(:,1:4))';S2=cov(data(:,5:8))';
$\mathrm{n} 1=$ length $(\mathrm{y} 11) ; \mathrm{n} 2=$ length $(\mathrm{y} 21)$;
$\mathrm{Spl}=(1 /(\mathrm{n} 1+\mathrm{n} 2-2)) *((\mathrm{n} 1-1) * \mathrm{~S} 1+(\mathrm{n} 2-1) * \mathrm{~S} 2)$;
$a=\operatorname{inv}(\mathrm{Spl}) *(\mathrm{~m} 1-\mathrm{m} 2)$,
Ans.
a =

$$
\begin{array}{r}
0.51042 \\
-0.20329 \\
0.46604 \\
-0.30967
\end{array}
$$

## 2.6: COMPUTATION OF $T^{2}$

If one has a program available with matrix manipulation capability, it is a simple matter to compute $T^{2}$ using (2.9). However, this approach is somewhat cumbersome for those not accustomed to the use of such a programming language, and many would prefer a more automated procedure. But very few general-purpose statistical programs provide for direct calculation of the two-sample $T^{2}$-statistic, perhaps because it is so easy to obtain from other procedures. We will discuss two types of widely available procedures that can be used to compute $T^{2}$.

### 2.6.1: Obtaining $T^{2}$ from a MANOVA Program

Multivariate analysis of variance (MANOVA) is discussed in Chapter 3, and the reader may wish to return to the present section after becoming familiar with that material. One-way MANOVA involves a comparison of mean vectors from several samples. Typically, the number of samples is three or more, but the procedure will also accommodate two samples. The two-sample $T^{2}$ test is thus a special case of MANOVA.

Four common test statistics are defined in Section 3.1: Wilks’ $\Lambda$, the Lawley-Hotelling $\mathrm{U}^{(s)}$, Pillai's $\mathrm{V}^{(s)}$, and Roy's largest root $\theta$. Without concerning ourselves here with how these are defined or calculated, we show how to use each to obtain the two-sample $T^{2}$ :

$$
\begin{align*}
& T^{2}=\left(n_{1}+n_{2}-2\right) \frac{1-\Lambda}{\Lambda}  \tag{2.16}\\
& T^{2}=\left(n_{1}+n_{2}-2\right) \mathrm{U}^{(s)}  \tag{2.17}\\
& T^{2}=\left(n_{1}+n_{2}-2\right) \frac{\mathrm{V}^{(s)}}{1-\mathrm{V}^{(s)}}  \tag{2.18}\\
& T^{2}=\left(n_{1}+n_{2}-2\right) \frac{\theta}{1-\theta} \tag{2.19}
\end{align*}
$$

(For the special case of two groups, $\mathrm{V}^{(s)}=\theta$.) These relationships are demonstrated in Section 3.1.7. If the MANOVA program gives eigenvectors of $\mathrm{E}^{-1} \mathrm{H}$ ( $\mathbf{E}$ and $\mathbf{H}$ are defined in Section 3.1.2), the eigenvector corresponding to the largest eigenvalue will be equal to (a constant multiple of) the discriminant function $\mathrm{S}_{\mathrm{pl}}^{-1}\left(\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}\right)$.

### 2.6.2: Obtaining $T^{2}$ from Multiple Regression

In this section, the $y$ 's become independent variables in a regression model. For each observation vector $\mathrm{y}_{1 i}$ and $\mathrm{y}_{2 i}$ in a two-sample $T^{2}$, define a "dummy" group variable as

$$
\begin{aligned}
w_{i} & =\frac{n_{2}}{n_{1}+n_{2}} \text { for each of } \mathrm{y}_{11}, \mathrm{y}_{12}, \cdots, \mathbf{y}_{1 n_{1}} \text { in sample } 1 \\
& =-\frac{n_{1}}{n_{1}+n_{2}} \text { for each of } \mathbf{y}_{21}, \mathbf{y}_{22}, \cdots, \mathbf{y}_{2 n_{2}} \text { in sample } 2
\end{aligned}
$$

Then $\bar{w}=0$ for all $n_{1}+n_{2}$ observations. The prediction equation for the regression of $w$ on the $y$ 's can be written as

$$
\hat{w}_{i}=b_{0}+b_{1} y_{i 1}+b_{2} y_{i 2}+\cdots+b_{p} y_{i p}
$$

where $i$ ranges over all $n_{1}+n_{2}$ observations and the least squares estimate $b_{0}$ is [see (7.15)]

$$
b_{0}=\bar{w}-b_{1} \bar{y}_{1}-b_{2} \bar{y}_{2}-\cdots-b_{p} \bar{y}_{p}
$$

Substituting this into the regression equation, we obtain

$$
\begin{aligned}
\hat{w}_{i} & =\bar{w}+b_{1}\left(y_{i 1}-\bar{y}_{1}\right)+b_{2}\left(y_{i 2}-\bar{y}_{2}\right)+\cdots+b_{p}\left(y_{i p}-\bar{y}_{p}\right) \\
& =b_{1}\left(y_{i 1}-\bar{y}_{1}\right)+b_{2}\left(y_{i 2}-\bar{y}_{2}\right)+\cdots+b_{p}\left(y_{i p}-\bar{y}_{p}\right)(\text { since } \bar{w}=0)
\end{aligned}
$$

Let $\mathrm{b}^{\prime}=\left(b_{1}, b_{2}, \cdots, b_{p}\right)$ be the vector of regression coefficients and $R^{2}$ be the squared multiple correlation. Then we have the following relationships:

$$
\begin{align*}
T^{2} & =\left(n_{1}+n_{2}-2\right) \frac{R^{2}}{1-R^{2}}  \tag{2.20}\\
\mathrm{a} & =\mathrm{S}_{\mathrm{pl}}^{-1}\left(\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}\right)=\frac{n_{1}+n_{2}}{n_{1} n_{2}}\left(n_{1}+n_{2}-2+T^{2}\right) \mathrm{b} \tag{2.21}
\end{align*}
$$

Thus with ordinary multiple regression, one can easily obtain $T^{2}$ and the discriminant function $\mathrm{S}_{\mathrm{pl}}^{-1}\left(\overline{\mathrm{y}}_{1}-\bar{y}_{2}\right)$. We simply define $w_{i}$ as above for each of the $n_{1}+n_{2}$ observations, regress the $w$ 's on the $y$ 's, and use the resulting $R^{2}$ in (2.20). For $\mathbf{b}$, delete the intercept from the regression coefficients for use in (2.21). Actually, since only the relative values of the elements of $\mathrm{a}=\mathrm{S}_{\mathrm{pl}}^{-1}\left(\overline{\mathrm{y}}_{1}-\overline{\mathrm{y}}_{2}\right)$ are of interest, it is not necessary to
convert from $\mathbf{b}$ to a in (2.21). We can use $\mathbf{b}$ directly or standardize the values $b_{1}, b_{2}, \cdots, b_{p}$ as in Section 5.5.

Example 2.6.2: We illustrate the regression approach to computation of $T^{2}$ using the psychological data in Table 2.3. We set $w=n_{2} /\left(n_{1}+n_{2}\right)$ $=32 / 64=1 / 2$ for each observation in group 1 (males) and equal to $-n_{1} /\left(n_{1}+n_{2}\right)=-1 / 2$ in the second group (females). When $w$ is regressed on the $64 y$ 's, we obtain

$$
\left(\begin{array}{l}
b_{0} \\
b_{1} \\
b_{2} \\
b_{3} \\
b_{4}
\end{array}\right)=\left(\begin{array}{c}
-0.75114 \\
0.05117 \\
-0.02038 \\
0.046721 \\
-0.031044
\end{array}\right) \quad, \quad R^{2}=0.61153
$$

By (2.20), $T^{2}=\left(n_{1}+n_{2}-2\right) \frac{R^{2}}{1-R^{2}}=(32+32-2) \frac{0.61153}{1-0.61153}=97.601$
as was obtained before in Example 2.4.2. Note that $\mathrm{b}^{\prime}=\left(b_{1}, b_{2}, b_{3}, b_{4}\right)=$ (.05117,-.02038, .046721,-.031044), with the intercept deleted, is proportional to the discriminant function coefficient vector a from Example 2.5, as we would expect from (2.21).

## Example 2.5[The program name mt6.m] $\quad$ Applications using MATLAB

 clc
## clear all

data=[15 $17241413141221 ; 17153226141214$ 26; 15142923121921 21;13 121016121310 16;20 172628112016 16;15 $2126211291418 ; 1513$ 2622101318 24;135 $22221081323 ; 147301712201923 ; 171530271110$ $1127 ; 1717262012182525 ; 1720282414181326 ; 1515292414102528 ;$ 1819322813168 14;18 18312714813 25;15 $14262113162328 ; 181733$ 26162126 26;10 141917141714 14;18 213029161615 23; 1821342613 1623 24;13 1730242616 21;16 161616141622 26;11 152523171722 28;16 132616161316 14;16 $13232115142026 ; 181834241210129 ; 1615$ $282714172423 ; 15162924131518$ 20; 18193223111618 28;18 1633237 719 18;17 20212112157 28; 19193028656 13]; y11=data(:,1); y12= data(:,2); y13=data(:,3); y14=data(:,4);y21=data(:,5);y22=data(:,6); y23=data(:,7); $\mathrm{y} 24=$ data $(:, 8) ; \mathrm{n} 1=$ length $(\mathrm{y} 11) ; \mathrm{n} 2=$ length $(\mathrm{y} 21) ; \mathrm{w} 1=\mathrm{n} 2 /(\mathrm{n} 1+\mathrm{n} 2) ; \mathrm{w} 2=-\mathrm{n} 1 /(\mathrm{n} 1+\mathrm{n} 2)$;
data1=[y11 y12 y13 y14];data2=[y21 y22 y23 y24]; w=[ones(size(y11))*w1; ones(size(y21))*w2]; x=[data1;data2]; n=n1+n2;x=[ones(n,1) x];b=xlw, W=x*b; $\mathrm{e}=\mathrm{w}-\mathrm{W} ; \mathrm{Rs}=\operatorname{corr}(\mathrm{w}, \mathrm{W})^{\wedge} 2, \mathrm{Ts}=(\mathrm{n} 1+\mathrm{n} 2-2) * \mathrm{Rs} /(1-\mathrm{Rs})$

Ans.

```
b =
        -0.75114
            0.05117
            -0.02038
            0.046721
            -0.031044
Rs =
            0.61153
Ts =
        97.601
```


## 2.7: PAIRED OBSERVATIONS TEST

As usual, we begin with the univariate case to set the stage for the multivariate presentation.

### 2.7.1: Univariate Case

Suppose two samples are not independent because there exists a natural pairing between the ith observation $y_{i}$ in the first sample and the ith observation $x_{i}$ in the second sample for all $i$, as, for example, when a treatment is applied twice to the same individual or when subjects are matched according to some criterion, such as IQ or family background. With such pairing, the samples are often referred to as paired observations or matched pairs. The two samples thus obtained are correlated, and the two-sample test statistic in (2.9) is not appropriate because the samples must be independent in order for (2.9) to have a $t$-distribution. [The two-sample test in (2.9) is somewhat robust to heterogeneity of variances and to lack of normality but not to dependence.] We reduce the two samples to one by working with the differences between the paired observations, as in the following layout for two treatments applied to the same subject:

Difference

| Pair Number | Treatment 1 | Treatment 2 | $d_{i}=y_{i}-x_{i}$ |
| :---: | :---: | :---: | :---: |
| 1 | $y_{1}$ | $x_{1}$ | $d_{1}$ |
| 2 | $y_{2}$ | $x_{2}$ | $d_{2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $y_{n}$ | $x_{n}$ | $d_{n}$ |

To obtain a $t$-test, it is not sufficient to assume individual normality for each of $y$ and $x$. To allow for the covariance between $y$ and $x$, we need the additional assumption that $y$ and $x$ have a bivariate normal distribution with

$$
\mu=\binom{\mu_{y}}{\mu_{x}} \quad, \quad \Sigma=\left(\begin{array}{ll}
\sigma_{y}^{2} & \sigma_{y x} \\
\sigma_{y x} & \sigma_{x}^{2}
\end{array}\right)
$$

where $d_{i}=y_{i}-x_{i}$ is $N\left(\mu_{y}-\mu_{x}, \sigma_{d}^{2}\right)$, where $\sigma_{d}^{2}=\sigma_{y}^{2}-2 \sigma_{y x}+\sigma_{x}^{2}$. From $d_{1}, d_{2}, \cdots, d_{n}$ we calculate

$$
\bar{d}=\frac{1}{n} \sum_{i=1}^{n} d_{i} \text { and } s_{d}^{2}=\frac{1}{n-1} \sum_{i=1}^{n}\left(d_{i}-\bar{d}\right)^{2}
$$

To test $H_{0}: \mu_{y}=\mu_{x}$, that is, $H_{0}: \mu_{d}=0$, we use the one-sample statistic

$$
\begin{equation*}
t=\frac{\bar{d}}{s_{d} / \sqrt{n}} \tag{2.22}
\end{equation*}
$$

which is distributed as $t_{n-1}$ if $H_{0}$ is true. We reject $H_{0}$ in favor of $H_{1}: \mu_{d} \neq 0$ if $|\mathrm{t}|>t_{\alpha / 2, n-1}$. It is not necessary to assume $\sigma_{y}^{2}=\sigma_{x}^{2}$ because there are no restrictions on $\Sigma$.

This test has only $n-1$ degrees of freedom compared with $2(n-1)$ for the two independent-sample $t$-test (2.8). In general, the pairing reduces the within-sample variation $s_{d}$ and thereby increases the power.

If we mistakenly treated the two samples as independent and used (2.8) with $n_{1}=n_{2}=n$, we would have

$$
t=\frac{\bar{y}-\bar{x}}{s_{\mathrm{pl}} \sqrt{2 / n}}=\frac{\bar{y}-\bar{x}}{\sqrt{2 s_{\mathrm{pl}}^{2} / n}}
$$

However,

$$
E\left(\frac{2 s_{\mathrm{pl}}^{2}}{n}\right)=2 E\left[\frac{(n-1) s_{y}^{2}+(n-1) s_{x}^{2}}{(n+n-2) n}\right]=\frac{\sigma_{y}^{2}+\sigma_{x}^{2}}{n}
$$

whereas $\operatorname{Var}(\bar{y}-\bar{x})=\left(\sigma_{y}^{2}+\sigma_{x}^{2}-2 \sigma_{y x}\right) / n$. Thus if the test statistic for independent samples (2.8) is used for paired data, it does not have a $t$ distribution and, in fact, underestimates the true average $t$-value
(assuming $H_{0}$ is false), since $\sigma_{y}^{2}+\sigma_{x}^{2}>\sigma_{y}^{2}+\sigma_{x}^{2}-2 \sigma_{y x}$ if $\sigma_{y x}>0$, which would be typical in this situation. One could therefore use

$$
\begin{equation*}
t=\frac{\bar{y}-\bar{x}}{\sqrt{\left(s_{y}^{2}+s_{x}^{2}-2 s_{y x}\right) / n}} \tag{2.23}
\end{equation*}
$$

but $t=\sqrt{n} \bar{d} / s_{d}$ in (2.22) is equal to it and somewhat simpler to use.

### 2.7.2: Multivariate Case

Here we assume the same natural pairing of sampling units as in the univariate case, but we measure $p$ variables on each sampling unit. Thus $y_{i}$ from the first sample is paired with $x_{i}$ from the second sample, $i=1,2, \ldots, n$. In terms of two treatments applied to each sampling unit, this situation is as follows:

| Pair Number | Treatment 1 | Treatment 2 | Difference <br> $\mathbf{d}_{i}=\mathbf{y}_{i}-\mathbf{x}_{i}$ |
| :---: | :---: | :---: | :---: |
| 1 | $\mathbf{y}_{1}$ | $\mathbf{x}_{1}$ | $\mathbf{d}_{1}$ |
| 2 | $\mathbf{y}_{2}$ | $\mathbf{x}_{2}$ | $\mathbf{d}_{2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $\mathbf{y}_{n}$ | $\mathbf{x}_{n}$ | $\mathbf{d}_{n}$ |

In Section 2.7.1, we made the assumption that $y$ and $x$ have a bivariate normal distribution, in which $y$ and $x$ are correlated. Here we assume $y$ and $\mathbf{x}$ are correlated and have a multivariate normal distribution:

$$
\binom{\mathrm{y}}{\mathrm{x}} \quad \text { is } \quad N_{2 p}\left[\binom{\mu_{y}}{\mu_{x}},\left(\begin{array}{cc}
\Sigma_{y y} & \Sigma_{y x} \\
\Sigma_{x y} & \Sigma_{x x}
\end{array}\right)\right]
$$

To test $H_{0}: \mu_{d}=0$, which is equivalent to $H_{0}: \mu_{y}=\mu_{x}$ since $\mu_{d}=E(\mathrm{y}-\mathrm{x})=$ $\mu_{y}-\mu_{x}$, we calculate

$$
\overline{\mathrm{d}}=\frac{1}{n} \sum_{i=1}^{n} \mathrm{~d}_{i} \text { and } \mathrm{S}_{d}=\frac{1}{n-1} \sum_{i=1}^{n}\left(\mathrm{~d}_{i}-\overline{\mathrm{d}}\right)\left(\mathrm{d}_{i}-\overline{\mathrm{d}}\right)^{\prime}
$$

We then have

$$
\begin{equation*}
T^{2}=\overline{\mathrm{d}}^{\prime}\left(\frac{\mathrm{S}_{d}}{n}\right)^{-1} \overline{\mathrm{~d}}=n \overline{\mathrm{~d}}^{\prime} \mathrm{S}_{d}^{-1} \overline{\mathrm{~d}} \tag{2.24}
\end{equation*}
$$

Under $H_{0}$, this paired comparison $T^{2}$-statistic is distributed as $T_{p, n-1}^{2}$. We reject $H_{0}$ if $T^{2}>T_{\alpha, p, n-1}^{2}$. Note that $\mathrm{S}_{d}$ estimates $\operatorname{Cov}(\mathrm{y}-\mathrm{x})=$

