Chapter 2
Multivariate Normal Distribution

In this chapter, we define univariate and multivariate normal distribution density functions and then we discuss tests of differences of means for multiple variables simultaneously across groups.

2.1 Univariate Normal Distribution

Just to refresh memory, in the case of a single random variable, the probability distribution or density function of that variable x is represented by Equation (2.1):

$$\Phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{2.1}$$

2.2 Bivariate Normal Distribution

The bivariate distribution represents the joint distribution of two random variables. The two random variables x_1 and x_2 are related to each other in the sense that they are not independent of each other. This dependence is reflected by the correlation ρ between the two variables x_1 and x_2. The density function for the two variables jointly is

$$\Phi(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right.\right.$$
$$\left.\left. -\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\} \tag{2.2}$$

This function can be represented graphically as in Fig. 2.1:

The *Isodensity contour* is defined as the set of points for which the values of x_1 and x_2 give the same value for the density function Φ. This contour is given by Equation (2.3) for a fixed value of C, which defines a constant probability:

$$\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} = C \tag{2.3}$$
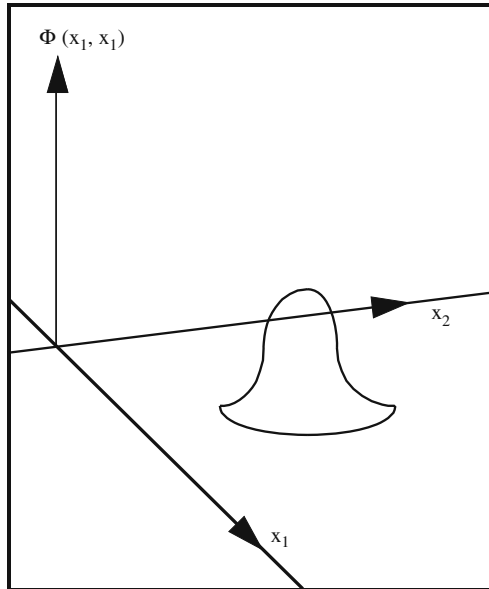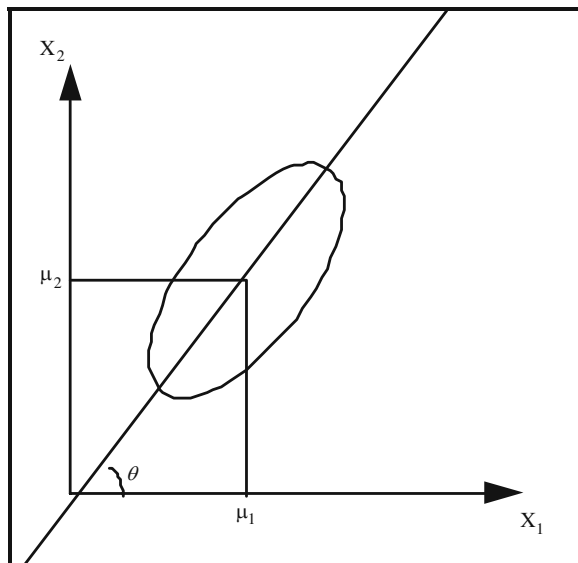
Fig. 2.1 The bivariate
normal distribution

Fig. 2.2 The locus of points
of the bivariate normal
distribution at a given density
level

Equation (2.3) defines an ellipse with centroid (μ_1, μ_2). This ellipse is the locus of points representing the combinations of the values of x_1 and x_2 with the same probability, as defined by the constant C (Fig. 2.2).

For various values of C, we get a family of concentric ellipses (at a different cut, i.e., cross section of the density surface with planes at various elevations) (see Fig. 2.3).

Fig. 2.3 Concentric ellipses
at various density levels
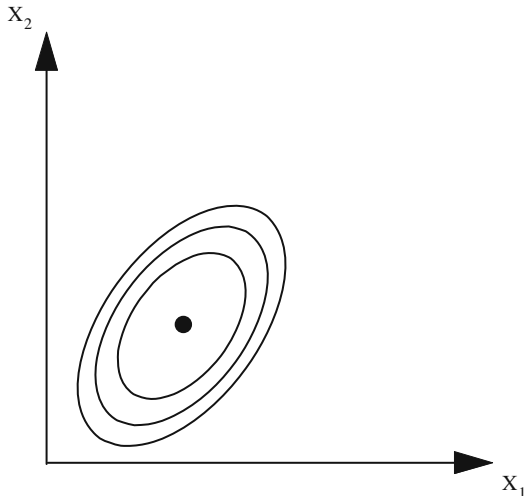
The angle θ depends only on the values of σ_1, σ_2, and ρ but is independent of C. The higher the correlation between x_1 and x_2, the steeper the line going through the origin with angle θ, i.e., the bigger the angle.

2.3 Generalization to Multivariate Case

Let us represent the bivariate distribution in matrix algebra notation in order to derive the generalized format for more than two random variables.

The covariance matrix of (x_1, x_2) can be written as

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{2.4}$$

The determinant of the matrix Σ is

$$|\Sigma| = \sigma_1^2\sigma_2^2\left(1 - \rho^2\right) \tag{2.5}$$

Equation (2.3) can now be re-written as

$$C = [x_1 - \mu_1, x_2 - \mu_2]\, \Sigma^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \tag{2.6}$$

where

$$\Sigma^{-1} = 1/\left[\sigma_1^2\sigma_2^2(1 - \rho^2)\right] \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} = \frac{1}{1 - \rho^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \tag{2.7}$$

Note that $\Sigma^{-1} = |\Sigma|^{-1} \times$ matrix of cofactors.
Let

$$\mathbf{X} = \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

then $\mathbf{X}'\Sigma^{-1}\mathbf{X} = \chi^2$, which is a quadratic form of the variables $\mathbf{x}$ and is, therefore, a chi-square variate.

Also, because $|\Sigma| = \sigma_1^2\sigma_2^2(1 - \rho^2)$, $|\Sigma|^{1/2} = \sigma_1\sigma_2\sqrt{(1 - \rho^2)}$, and consequently,

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} = (2\pi)^{-1}|\Sigma|^{-1/2} \tag{2.8}$$

the bivariate distribution function can be now expressed in matrix notation as

$$\Phi(x_1, x_2) = (2\pi)^{-1}|\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{X}'\Sigma^{-1}\mathbf{X}} \tag{2.9}$$

Now, more generally with p random variables $(x_1, x_2, \ldots, x_p)$, let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}; \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

The density function is

$$\Phi(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma|^{-\frac{1}{2}} e^{\left[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right]} \tag{2.10}$$

For a fixed value of the density Φ, an ellipsoid is described. Let $\mathbf{X} = \mathbf{x} - \mu$. The inequality $\mathbf{X}'\Sigma^{-1}\mathbf{X} \le \chi^2$ defines any point within the ellipsoid.

2.4 Tests About Means

2.4.1 Sampling Distribution of Sample Centroids

2.4.1.1 Univariate Distribution

A random variable is normally distributed with mean μ and variance σ^2:

$$x \sim N\left(\mu, \sigma^2\right) \tag{2.11}$$

After n independent draws, the mean is randomly distributed with mean μ and variance σ^2/n:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{2.12}$$

2.4.1.2 Multivariate Distribution

In the multivariate case with p random variables, where $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, $\mathbf{x}$ is normally distributed following the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ:

$$\mathbf{x} \sim N(\mu, \Sigma) \tag{2.13}$$

The mean vector for the sample of size n is denoted by

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

This sample mean vector is normally distributed with a multivariate normal distribution with mean μ and covariance Σ/n:

$$\bar{\mathbf{x}} \sim N\left(\mu, \frac{\Sigma}{n}\right) \tag{2.14}$$

2.4.2 Significance Test: One-Sample Problem

2.4.2.1 Univariate Test

The univariate test is illustrated in the following example. Let us test the hypothesis that the mean is 150 (i.e., $\mu_0 = 150$) with the following information:

$$\sigma^2 = 256; \ n = 64; \ \bar{x} = 154$$

Then, the z score can be computed as

$$z = \frac{154 - 150}{\sqrt{256/64}} = \frac{4}{16/8} = 2$$

At $\alpha = 0.05$ (95% confidence interval), $z = 1.96$, as obtained from a normal distribution table. Therefore, the hypothesis is rejected. The confidence interval is

$$\left[154 - 1.96 \times \frac{12}{6}, \ 154 + 1.96 \times \frac{12}{6}\right] = [150.08, \ 157.92]$$

This interval excludes 150. The hypothesis that $\mu_0 = 150$ is rejected. If the variance σ had been unknown, the t statistic would have been used:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{2.15}$$

where s is the observed sample standard deviation.

2.4.2.2 Multivariate Test with Known Σ

Let us take an example with two random variables:

$$\Sigma = \begin{bmatrix} 25 & 10 \\ 10 & 16 \end{bmatrix} \qquad n = 36$$

$$\bar{\mathbf{x}} = \begin{bmatrix} 20.3 \\ 12.6 \end{bmatrix}$$

The hypothesis is now about the mean values stated in terms of the two variables jointly:

$$H: \quad \mu_0 = \begin{bmatrix} 20 \\ 15 \end{bmatrix}$$

At the alpha level of 0.05, the value of the density function can be written as below, which follows a chi-squared distribution at the specified significance level α:

$$n\,(\mu_\circ - \bar{\mathbf{x}})'\,\Sigma^{-1}\,(\mu_0 - \bar{\mathbf{x}}) \sim \chi_p^2\,(\alpha) \tag{2.16}$$

Computing the value of the statistics,

$$|\Sigma| = 25 \times 16 - 10 \times 10 = 300$$

$$\Sigma^{-1} = \frac{1}{300} \begin{bmatrix} 16 & -10 \\ -10 & 25 \end{bmatrix}$$

$$\chi^2 = 36 \times \frac{1}{300}\,(20 - 20.3, \; 15 - 12.6) \begin{bmatrix} 16 & -10 \\ -10 & 25 \end{bmatrix} \begin{bmatrix} 20 - 20.3 \\ 15 - 12.6 \end{bmatrix} = 15.72$$

The critical value at an alpha value of 0.05 with two degrees of freedom is provided by tables:

$$\chi_{p=2}^2\,(\alpha = 0.05) = 5.991$$

The observed value is greater than the critical value. Therefore, the hypothesis that $\mu = \begin{bmatrix} 20 \\ 15 \end{bmatrix}$ is rejected.

2.4.2.3 Multivariate Test with Unknown Σ

Just as in the univariate case, Σ is replaced with the sample value $\mathbf{S}/(n-1)$, where $\mathbf{S}$ is the sum-of-squares-and-cross-products (SSCP) matrix, which provides

an unbiased estimate of the covariance matrix. The following statistics are then used to test the hypothesis:

$$\text{Hotelling: } T^2 = n\,(n-1)\,(\bar{\mathbf{x}} - \mu_0)\,{}'\mathbf{S}^{-1}\,(\bar{\mathbf{x}} - \mu_0) \qquad (2.17)$$

where, if

$$\mathbf{X}^d_{n \times p} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots \\ \vdots & \vdots & \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \cdots \end{bmatrix}$$

$$\mathbf{S} = \mathbf{X}^{d'}\mathbf{X}^d$$

Hotelling showed that

$$\frac{n-p}{(n-1)\,p}\,T^2 \sim F^p_{n-p} \qquad (2.18)$$

Replacing T^2 by its expression given above

$$\frac{n\,(n-p)}{p}\,(\bar{\mathbf{x}} - \mu_0)'\,\mathbf{S}^{-1}\,(\bar{\mathbf{x}} - \mu_0) \sim F^p_{n-p} \qquad (2.19)$$

Consequently, the test is performed by computing the expression above and comparing its value with the critical value obtained in an F table with p and $n-p$ degrees of freedom.

2.4.3 Significance Test: Two-Sample Problem

2.4.3.1 Univariate Test

Let us define $\bar{x}_1$ and $\bar{x}_2$ as the means of a variable on two unrelated samples. The test for the significance of the difference between the two means is given by

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{or} \quad t^2 = \frac{(\bar{x}_1 - \bar{x}_2)^2}{s^2\left(\frac{n_1 + n_2}{n_1 n_2}\right)} \qquad (2.20)$$

where

$$s = \frac{\sqrt{(n_1 - 1)\frac{\sum_i x_{1i}^2}{n_1 - 1} + (n_2 - 1)\frac{\sum_i x_{2i}^2}{n_2 - 1}}}{(n_1 - 1) + (n_2 - 1)} = \sqrt{\frac{\sum_i x_{1i}^2 + \sum_i x_{2i}^2}{n_1 + n_2 - 2}} \qquad (2.21)$$

s^2 is the pooled within groups variance. It is an estimate of the assumed common variance σ^2 of the two populations.

2.4.3.2 Multivariate Test

Let $\bar{\mathbf{x}}^{(1)}$ be the mean vector in sample $1 = \begin{bmatrix} \bar{x}_1^{(1)} \\ \bar{x}_2^{(1)} \\ \vdots \\ \bar{x}_p^{(1)} \end{bmatrix}$ and similarly for sample 2.

We need to test the significance of the difference between $\bar{\mathbf{x}}^{(1)}$ and $\bar{\mathbf{x}}^{(2)}$. We will consider first the case where the covariance matrix, which is assumed to be the same in the two samples, is known. Then we will consider the case where an estimate of the covariance matrix needs to be used.

Σ Is Known (The Same in the Two Samples)

In this case, the difference between the two group means is normally distributed with a multivariate normal distribution:

$$\left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} \right) \sim N \left(\mu_1 - \mu_2, \Sigma \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right) \qquad (2.22)$$

The computations for testing the significance of the differences are similar to those in Section 2.4.2.2 using the chi-square test.

Σ Is Unknown

If the covariance matrix is not known, it is estimated using the covariance matrices within each group but pooled.

Let $\mathbf{W}$ be the within-groups SSCP (sum of squares cross products) matrix. This matrix is computed from the matrix of deviations from the means on all p variables for each of n_k observations (individuals). For each group k,

$$\underset{n_k \times p}{\mathbf{X}}^{d(k)} = \begin{bmatrix} x_{11}^{(k)} - \bar{x}_1^{(k)} & x_{21}^{(k)} - \bar{x}_2^{(k)} & \cdots \\ x_{12}^{(k)} - \bar{x}_1^{(k)} & x_{22}^{(k)} - \bar{x}_2^{(k)} & \cdots \\ \vdots & \vdots & \\ x_{1n_k}^{(k)} - \bar{x}_1^{(k)} & x_{2n_k}^{(k)} - \bar{x}_2^{(k)} & \cdots \end{bmatrix} \qquad (2.23)$$

For each of the two groups (each k), the SSCP matrix can be derived:

$$\mathbf{S}_k = \mathbf{X}_{p\times n_k}^{d(k)'} \mathbf{X}_{n_k\times p}^{d(k)} \tag{2.24}$$

The pooled SSCP matrix for the more general case of K groups is simply:

$$\underset{p\times p}{\mathbf{W}} = \sum_{k=1}^{K} \underset{p\times p}{\mathbf{S}_k} \tag{2.25}$$

In the case of two groups, K is simply equal to 2.

Then, we can apply Hotelling's T, just as in Section 2.4.2.3, where the proper degrees of freedom depending on the number of observations in each group (n_k) are applied.

$$T^2 = \left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right)' \left[\frac{\mathbf{W}}{n_1 + n_2 - 2}\frac{n_1 + n_2}{n_1 n_2}\right]^{-1} \left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right) \tag{2.26}$$

$$= \frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2} \left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right)' \mathbf{W}^{-1} \left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right) \tag{2.27}$$

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}T^2 \sim F_{n_1 + n_2 - p - 1}^p \tag{2.28}$$

2.4.4 Significance Test: K-Sample Problem

As in the case of two samples, the null hypothesis is that the mean vectors across the K groups are the same and the alternative hypothesis is that they are different.

Let us define Wilk's likelihood-ratio criterion:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} \tag{2.29}$$

where $\mathbf{T}$ = total SSCP matrix, $\mathbf{W}$ = within-groups SSCP matrix.

$\mathbf{W}$ is defined as in Equation (2.25). The total SSCP matrix is the sum of squared cross products applied to the deviations from the grand means (i.e., the overall mean across the total sample with the observations of all the groups for each variable). Therefore, let the mean centered data for group k be noted as

$$\underset{n_k\times p}{\mathbf{X}^{d*(k)}} = \begin{bmatrix} x_{11}^{(k)} - \bar{x}_1 & x_{21}^{(k)} - \bar{x}_2 & \cdots \\ x_{12}^{(k)} - \bar{x}_1 & x_{22}^{(k)} - \bar{x}_2 & \cdots \\ \vdots & \vdots & \\ x_{1n_k}^{(k)} - \bar{x}_1 & x_{2n_k}^{(k)} - \bar{x}_2 & \cdots \end{bmatrix} \tag{2.30}$$

where $\bar{x}_j$ is the overall mean of the j's variate.

Bringing the centered data for all the groups in the same data matrix leads to

$$\underset{nxp}{\mathbf{X}^{d*}} = \begin{bmatrix} \mathbf{X}^{d*(1)} \\ \mathbf{X}^{d*(2)} \\ \vdots \\ \mathbf{X}^{d*(K)} \end{bmatrix} \tag{2.31}$$

The total SSCP matrix T is then defined as

$$\underset{p\times p}{\mathbf{T}} = \underset{p\times n}{\mathbf{X}^{d*\prime}} \underset{n\times p}{\mathbf{X}^{d*}} \tag{2.32}$$

Intuitively, if we reduce the space to a single variate so that we are only dealing with variances and no covariances, Wilk's lambda is the ratio of the pooled within-group variance to the total variance If the group means are the same, the variances are equal and the ratio equals one. As the group means differ, the total variance becomes larger than the pooled within-group variance. Consequently, the ratio lambda becomes smaller. Because of the existence of more than one variate, which implies more than one variance and covariances, the within SSCP and Total SSCP matrices need to be reduced to a scalar in order to derive a scalar ratio. This is the role of the determinants. However, the interpretation remains the same as for the univariate case.

It should be noted that Wilk's Λ can be expressed as a function of the Eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ where $\mathbf{B}$ is the between-group covariance matrix (Eigenvalues are explained in the next chapter). From the definition of Λ in Equation (2.29), it follows that

$$\frac{1}{\Lambda} = \frac{|\mathbf{T}|}{|\mathbf{W}|} = \left|\mathbf{W}^{-1}\mathbf{T}\right| = \left|\mathbf{W}^{-1}(\mathbf{W}+\mathbf{B})\right| = \left|\mathbf{I}+\mathbf{W}^{-1}\mathbf{B}\right| = \prod_{i=1}^{K}(1+\lambda_i) \tag{2.33}$$

and consequently

$$\Lambda = \frac{1}{\prod_{i=1}^{K}(1+\lambda_i)} = \prod_{i=1}^{K}\frac{1}{(1+\lambda_i)} \tag{2.34}$$

Also, it follows that

$$\mathrm{Ln}\Lambda = \mathrm{Ln}\frac{1}{\prod_{i=1}^{K}(1+\lambda_i)} = -\sum_{i=1}^{K}(1+\lambda_i) \tag{2.35}$$

When Wilk's Λ approaches 1, we showed that it means that the difference in means is negligible. This is the case when $\mathrm{Ln}\Lambda$ approaches 0. However, when Λ approaches 0 or $\mathrm{Ln}\Lambda$ approaches 1, it means that the difference is large. Therefore, a large value of $\mathrm{Ln}\Lambda$ (i.e., close to 0) is an indication of the significance of the difference between the means.

Based on Wilk's lambda, we present two statistical tests: Bartlett's V and Rao's R.

Let n = total sample size across samples, p = number of variables, and K = number of groups (number of samples).

Bartlett's V is approximately distributed as a chi-square when $n - 1 - (p + K)/2$ is large:

$$V = -\left[n - 1 - (p + K)/2\right] Ln\Lambda \sim \chi^2_{p(K-1)} \tag{2.36}$$

Bartlett's V is relatively easy to calculate and can be used when $n - 1 - (p + K)/2$ is large.

Another test can be applied, as Rao's R is distributed approximately as an F variate. It is calculated as follows:

$$R = \frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \frac{ms - p(K-1)/2 + 1}{p(K-1)} \approx F^{v_1 = p(K-1)}_{v_2 = ms - p(K-1)/2 + 1} \tag{2.37}$$

where

$$m = n - 1 - (p + K)/2$$

$$s = \sqrt{\frac{p^2(K-1)^2 - 4}{p^2 + (K-1)^2 - 5}}$$

2.5 Examples Using SAS

2.5.1 Test of the Difference Between Two Mean Vectors – One-Sample Problem

In this example, the file "MKT_DATA" contains data about the market share of a brand over seven periods, as well as the percentage of distribution coverage and the price of the brand. These data correspond to one market, Norway. The question is to know whether the market share, distribution coverage, and prices are similar or different from the data of that same brand for the rest of Europe, i.e., with values of market share, distribution coverage, and price, respectively of 0.17, 32.28, and 1.39. The data are shown below in Table 2.1:

Table 2.1 Data example for the analysis of three variables

PERIOD	M_SHARE	DIST	PRICE
1	0.038	11	0.98
2	0.044	11	1.08
3	0.039	9	1.13
4	0.03	9	1.31
5	0.036	14	1.36
6	0.051	14	1.38
7	0.044	9	1.34

```
/* *********** Example2-1.sas ************* */
OPTIONS LS=80;
DATA work;
INFILE
"C:\SAMD2\Chapter2\Examples\Mkt_Data.csv"
dlm = ',' firstobs=2;
INPUT PERIOD M_SHARE DIST PRICE;
data work;
        set work (drop = period) ;
run;
/* Multivariate Test with Unknown Sigma */
proc iml;
print " Multivariate Test with Unknown Sigma " ;
print "-----------------------------------" ;
use work;                       /* Specifying the matrix with raw market data for Norway */
read all var {M_Share Dist Price} into Mkt_Data;
start SSCP;                      /* SUBROUTINE for calculation of the SSCP matrix */
        n=nrow(x);                      /* Number of rows */
        mean=x[+,]/n;                   /* Column means    */
        x=x-repeat(mean,n,1);           /* Variances       */
        sscp = x`*x;                    /* SSCP matrix     */
finish sscp;                            /* END SUBROUTINE */
x=Mkt_Data;                             /* Definition of the data matrix */
p=ncol(Mkt_Data);
run sscp;                               /* Execution of the SUBROUTINE    */
print SSCP  n p;

Xbar = mean;                            /* Definition of the mean vector */
m_o = { 0.17 32.28 1.39 };              /* Myu zero: the mean vector for Europe */

dX = Xbar - m_o;                        /* Matrix of deviations */
dXt = dX`;                              /* Calculation of the transpose of dX */

print m_o;
print Xbar;
print dX;

sscp_1 = inv(sscp);             /* Calculation of the inverse of SSCP matrix */

T_sq = n*(n-1)*dX*sscp_1*dXt;           /* Calculation of the T_square    */
F    = T_sq*(n-p)/((n-1)*p);            /* Calculation of the F statistic */

Df_num = p;
Df_den = n-p ;
F_crit = finv(.95,df_num,df_den);    /* Critical F for .05 for df_num, df_den */
Print F F_crit;
quit;
```

Fig. 2.4 SAS input to perform the test of a mean vector (examp2-1.sas)

The SAS file showing the SAS code to compute the necessary statistics is shown below in Fig. 2.4. The first lines correspond to the basic SAS instructions to read the data from the file. Here, the data file was saved as a text file from Excel. Consequently, the values in the file corresponding to different data points are separated by commas. This is indicated as the delimiter ("dlm"). Also, the data (first observation) starts on line 2 because the first line is used for the names of the variables (as illustrated in Table 2.1). The variable called period is dropped so that only the three variables needed for the analysis are kept in the SAS working data set. The procedure IML is used to perform matrix algebra computations.

This file could easily be used for the analysis of different databases. Obviously, it would be necessary to adapt some of the instructions, especially the file name and path and the variables. Within the IML subroutine, only two things would need to be changed: (1) the variables used for the analysis and (2) the values for the null hypothesis (m_o).

Fig. 2.5 SAS output of
analysis defined in Fig. 2.4
(examp2-1.lst)

```
                 Multivariate Test with Unknown Sigma

        -------------------------------------

    SSCP                                         N        P

0.0002734      0.035 0.0007786        7        3
    0.035         30      0.66
0.0007786       0.66 0.1527714


                        M_O

            0.17     32.28       1.39


                       XBAR

    0.0402857            11 1.2257143


                        DX

    -0.129714      -21.28 -0.164286


                    F      F_CRIT

                588.72944 6.5913821
```

The results are printed in the output file shown below in Fig. 2.5:

The critical *F* statistic with three and four degrees of freedom at the 0.05 confidence level is 6.591, while the computed value is 588.7, indicating that the hypothesis of no difference is rejected.

2.5.2 Test of the Difference Between Several Mean Vectors – K-Sample Problem

The next example considers similar data for three different countries (Belgium, France, and England) for seven periods, as shown in Table 2.2. The question is to know whether the mean vectors are the same for the three countries or not.

We first present an analysis that shows the matrix computations following precisely the equations presented in Section 2.4.4. These involve the same matrix manipulations in SAS as in the prior example, using the IML procedure in SAS. Then we present the MANOVA analysis proposed by SAS using the GLM procedure. The readers who want to skip the detailed calculations can go directly to the SAS GLM procedure.

The SAS file which derived the computations for the test statistics is shown in Fig. 2.6.

The results are shown in the SAS output on Fig. 2.7.

These results indicate that the Bartlett's *V* statistic of 82.54 is larger than the critical chi-square with six degrees of freedom at the 0.05 confidence level (which is 12.59). Consequently, the hypothesis that the mean vectors are the same is rejected.

Table 2.2 Data example for three variables in three countries (groups)

CNTRYNO	CNTRY	PERIOD	M_SHARE	DIST	PRICE
1	BELG	1	0.223	61	1.53
1	BELG	2	0.22	69	1.53
1	BELG	3	0.227	69	1.58
1	BELG	4	0.212	67	1.58
1	BELG	5	0.172	64	1.58
1	BELG	6	0.168	64	1.53
1	BELG	7	0.179	62	1.69
2	FRAN	1	0.038	11	0.98
2	FRAN	2	0.044	11	1.08
2	FRAN	3	0.039	9	1.13
2	FRAN	4	0.03	9	1.31
2	FRAN	5	0.036	14	1.36
2	FRAN	6	0.051	14	1.38
2	FRAN	7	0.044	9	1.34
3	UKIN	1	0.031	3	1.43
3	UKIN	2	0.038	3	1.43
3	UKIN	3	0.042	3	1.3
3	UKIN	4	0.037	3	1.43
3	UKIN	5	0.031	13	1.36
3	UKIN	6	0.031	14	1.49
3	UKIN	7	0.036	14	1.56

The same conclusion could be derived from the Rao's R statistic with its value of 55.10, which is larger than the corresponding F value with 6 and 32 degrees of freedom, which is 2.399.

The first lines of SAS code in Fig. 2.8 read the data file in the same manner as in the prior examples. However, the code that follows is much simpler as the procedure automatically performs the MANOVA tests. For that analysis, the general procedure of the General Linear Model is called with the statement "proc glm". The class statement indicates that the variable that follows (here "CNTRY") is a discrete (nominal scaled) variable. This is the variable used to determine the K groups. K is calculated automatically according to the different values contained in the variable. The model statement shows the list of the variates for which the means will be compared on the left-hand side of the equal sign. The variable on the right-hand side is the group variable. The GLM procedure is in fact a regression where the dependent variables are regressed on the dummy variables automatically created by SAS reflecting the various values of the grouping variable. The optional parameter "nouni" after the slash indicates that the univariate tests should not be performed (and consequently their corresponding output will not be shown). Finally, the last line of code necessarily indicates that the MANOVA test concerns the differences across the grouping variable, CNTRY.

The output shown in Fig. 2.9 provides the same information as shown in Fig. 2.7. Wilk's Lambda has the same value of 0.007787. In addition, several other tests are provided for its significance, leading to the same conclusion that the differences in

```
/*  ***************** Examp2-2.sas ******************  */
OPTIONS LS=80;
DATA work;
INFILE
"C:\SAMD2\CHAPTER2\EXAMPLES\Mkt_Dt_K.csv"
dlm = ',' firstobs=2;
INPUT CNTRYNO CNTRY $ PERIOD M_SHARE DIST PRICE;
data work;
        set work (drop = cntry period) ;
proc print;
proc freq;
tables cntryno / out = Nk_out (keep = count);
run;
/* Significance Test: K-Sample Problem */
proc iml;
reset center;
print " Multivariate Significance Test: K-Sample Problem " ;
print "-------------------------------------------------" ;
use work ;        /* Specifying the matrix with raw data */
read all var { CNTRYNO M_SHARE DIST PRICE}  into Mkt_Data;
use Nk_out;
read all var {count} into Nk_new;
                        /* Number of observations within each group       */
n_tot = nrow(Mkt_Data);
K=max(Mkt_Data[,1]);            /* Number of groups (samples) */
p=ncol(Mkt_Data)-1;             /* Number of variables */
print n_tot "    " K "   " p;
start SSCP;                     /* SUBROUTINE for calculation of the SSCP matrix
*/
        n=nrow(x) ;
        mean=x[+,]/n;           /* Column means (mean vector)
*/
        x=x-repeat(mean,n,1);   /* Matrix of variances
*/
        SSCP = x`*x;            /* SSCP matrix
*/
print i "     "  mean;
finish SSCP;                    /* END SUBROUTINE
*/
S = J(p,p,0);                   /* Definition of a p x p square matrix with zeros
*/
do i = 1 to K;
if i = 1 then a = 1;
else
a=1+(i-1)*nk_new[i-1];
b=a+nk_new[i]-1;
x = Mkt_Data[a:b,2:4];
run SSCP;                       /* Execution of the SUBROUTINE for each group
*/
S = S + SSCP;                   /* Accumulation of the sum of SSCP matrices
*/
end;                            /* in order to calculate W (within-the-groups SSCP)
*/
W = S; DetW = Det(W);
print W "     " DetW;
x=Mkt_Data[,2:4];               /* Definition of the data matrix (dropping the first
column: CNTRYNO) */
run SSCP;                       /* Execution of the SUBROUTINE for total data
*/
T=SSCP;
DetT = Det(T);
print T "     " DetT;
Lmbd = Det(W) / Det(T);
m = n_tot-1-(p+K) / 2;
reset noname fw=5 nocenter;
print "Lambda =" Lmbd [format=10.6];
print "m =" m [format=2.0]
                " Use Bartlett's V for large m's and Rao's R otherwise " ;
V = -m*Log(Lmbd);
s = sqrt((p*p*(K-1)**2-4)/(p*p+(K-1)**2-5));
R = (1-Lmbd**(1/s))*(m*s-p*(K-1)/2 + 1)/(Lmbd**(1/s)*p*(K-1));
Df_num = p*(K-1);                       Df_den = m*s-Df_num/2 +1 ;
Chi_crit = CINV(0.95,Df_num);           F_crit = finv(.95,df_num,df_den);
print "Bartlett's V = " V [format=9.6] "   DF =" DF_num [format=2.0]   ;
print "     Chi_crit =" Chi_crit [format=9.6];
print "Rao's R ="  R [format=9.6]
                        "    DF_NUM =" Df_num [format=2.0]
                        "    DF_DEN =" Df_den [format=2.0] ;
print "     F_crit =" F_crit [format=9.6];
quit;
```

Fig. 2.6 SAS input to perform a test of difference in mean vectors across *K* groups (examp2-2.sas)

```
              Multivariate Significance Test: K-Sample Problem
         ---------------------------------------------------
                    N_TOT                 K                 P
                     21                   3                 3

               I                MEAN
               1           0.2001429 65.142857 1.5742857

               I                MEAN
               2           0.0402857           11 1.2257143

               I                MEAN
               3           0.0351429 7.5714286 1.4285714

               W                                      DETW
        0.0044351 0.2002857 -0.002814           0.246783
        0.2002857 288.57143 1.8214286
        -0.002814 1.8214286 0.2144286

               I                MEAN
               4           0.0918571 27.904762 1.4095238

               T                                      DETT
        0.1276486 42.601714 0.1808686           31.691145
        42.601714  14889.81 63.809048
        0.1808686 63.809048 0.6434952

  Lambda =    0.007787

  m = 17   Use Bartlett's V for large m's and Rao's R otherwise

  Bartlett's V =  82.539814      DF =  6
        Chi_crit = 12.591587

  Rao's R = 55.104665     DF_NUM =  6     DF_DEN = 32
        F_crit =   2.399080
```

Fig. 2.7 SAS output of test of difference across *K* groups (examp2-2.lst)

```
/*  ***************** Examp2-3-Manovasas.sas ****************  */

OPTIONS LS=80;
DATA work;
INFILE
""C:\SAMD2\CHAPTER2\EXAMPLES\Mkt_Dt_K.csv"
dlm = ',' firstobs=2;
INPUT CNTRYNO CNTRY $ PERIOD M_SHARE DIST PRICE;


/* Chapter 2, IV.4 Significance Test: K-Sample Problem */
proc glm;
      class CNTRY;
      model M_SHARE DIST PRICE=CNTRY /nouni;
      manova h = CNTRY/ printe;
run;

quit;
```

Fig. 2.8 SAS input for MANOVA test of mean differences across *K* groups (examp2-3.sas)

```
                          The GLM Procedure

                      Class Level Information

               Class        Levels    Values

               CNTRY           3       BELG FRAN UKIN


               Number of Observations Read        21
               Number of Observations Used        21

                  Multivariate Analysis of Variance

                     E = Error SSCP Matrix


                   M_SHARE              DIST                PRICE

      M_SHARE     0.0044351429      0.2002857143       -0.002814286
      DIST        0.2002857143      288.57142857        1.8214285714
      PRICE      -0.002814286       1.8214285714        0.2144285714


  Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

            DF = 18       M_SHARE           DIST           PRICE

          M_SHARE       1.000000         0.177039       -0.091258
                                          0.4684          0.7102

          DIST          0.177039         1.000000        0.231550
                         0.4684                           0.3402

          PRICE        -0.091258         0.231550        1.000000
                         0.7102           0.3402


        Characteristic Roots and Vectors of: E Inverse * H, where
                H = Type III SSCP Matrix for CNTRY
                     E = Error SSCP Matrix

    Characteristic              Characteristic Vector  V'EV=1
         Root      Percent       M_SHARE           DIST           PRICE

     67.2013787     98.70       7.5885004       0.0457830       0.0045113
      0.8829099      1.30       3.7773797      -0.0204742       2.2231712
      0.0000000      0.00     -12.8623871       0.0361429       0.2847771



              MANOVA Test Criteria and F Approximations for
               the Hypothesis of No Overall CNTRY Effect
                H = Type III SSCP Matrix for CNTRY
                     E = Error SSCP Matrix

                    S=2     M=0     N=7

Statistic                     Value    F Value   Num DF   Den DF    Pr > F

Wilks' Lambda             0.00778713     55.10       6       32     <.0001
Pillai's Trace            1.45424468     15.10       6       34     <.0001
Hotelling-Lawley Trace   68.08428858    176.86       6   19.652     <.0001
Roy's Greatest Root      67.20137868    380.81       3       17     <.0001

          NOTE: F Statistic for Roy's Greatest Root is an upper bound.
              NOTE: F Statistic for Wilks' Lambda is exact.
```

Fig. 2.9 SAS output for MANOVA test of mean differences across *K* groups (examp2-3.lst)

```
/******************************************************************************
 Assign2.sas
 Creation of additional data files for Chapter2 assignments.
 ******************************************************************************/
option ls=120 ;
/*----------------------------------------------------------------------------
 Creating the dataset PANEL by reading data from c:\...\panel.csv
 ----------------------------------------------------------------------------*/
data panel;
  infile 'C:\SAMD2\Chapter2\Assignments\panel.csv' firstobs=2   dlm = ',' ;
  input period    segment  segsize  ideal1-ideal3
        brand $  adv_pct   aware    intent shop1-shop3
        perc1-perc3 dev1-dev3 share ;
run;
proc sort data=panel;
        by period brand;
run;
/*----------------------------------------------------------------------------
 Creating the dataset INDUP by reading data from c:\...\indup.csv
 ----------------------------------------------------------------------------*/
data indup;
  infile 'C:\SAMD2\Chapter2\Assignments\indup.csv' firstobs=2   dlm = ',' ;
  input  period  firm  brand $ price  advert
         char1-char5 salmen1-salmen3
         cost dist1-dist3 usales  dsales  ushare dshare  adshare  relprice ;
run;
proc sort data =indup;
        by period brand;
run;
/*----------------------------------------------------------------------------
 Merging PANEL and INDUP into ECON
 ----------------------------------------------------------------------------*/
data econ;
        merge panel indup;
        by period brand;
if segment<5 then delete;
run;
proc means noprint;
var intent share ;
output out = econmean mean=IntMean ShrMean;
run;
/*----------------------------------------------------------------------------
 Writing EconMean to a CSV file (easily opened by Excel)
 ----------------------------------------------------------------------------*/
data _NULL_;
  set EconMean (keep = IntMean ShrMean);
  by IntMean ;
  TAB = ',' ;
  FN = "C:\SAMD2\CHAPTER2\ASSIGNMENTS\Mean1grp.CSV";
  file PLOTFILE filevar=FN;
  if ( FIRST.IntMean ) then
  do;
    put "IntMean" TAB "ShrMean" ;
  end;
    put  IntMean  TAB  ShrMean  ;
run;
/*----------------------------------------------------------------------------
 Creating a new dataset EconNew with selected variables from ECON
 ----------------------------------------------------------------------------*/
data EconNew;
set  Econ   ;
keep segment period brand intent share ;
```

Fig. 2.10 Example of SAS file for reading data sets INDUP and PANEL and creating new data files (assign2.sas)

means are significant. In addition to the expression of Wilk's lambda as a function of the Eigenvalues of $\mathbf{W^{-1}B}$, three other measures are provided in the SAS output.

Pillai′s Trace is defined as $\displaystyle\sum_{i=1}^{K} \frac{\lambda_i}{1+\lambda_i}$

Hotelling–Lawley Trace is simply the sum of the Eigenvalues: $\displaystyle\sum_{i=1}^{K} \lambda_i$

Roy's Greatest Root is the ratio $\frac{\lambda_{\max}}{1+\lambda_{\max}}$

These tests tend to be consistent, but the numbers are different. As noted in the SAS output, Roy's Greatest Root is an upper bound to the statistic.

2.6 Assignment

In order to practice with these analyses, you will need to use the databases INDUP and PANEL described in Appendix C. These databases provide market share and marketing mix variables for a number of brands competing in five market segments. You can test the following hypotheses:

1. The market behavioral responses of a given brand (e.g., awareness, perceptions, or purchase intentions) are different across segments,
2. The marketing strategy (i.e., the values of the marketing mix variables) of selected brands is different (perhaps corresponding to different strategic groups).

Figure 2.10 shows how to read the data within a SAS file and how to create new files with a subset of the data saved in a format, which can be read easily using the examples provided throughout this chapter. Use the model described in the examples above and adapt them to the database to perform these tests.

```
where brand = 'salt';
run;
proc sort ;
by Brand Segment Period ;
run;
/*-----------------------------------------------------------------------------
 Writing EconNew to a CSV file (easily opened by Excel)
-----------------------------------------------------------------------------*/
data _NULL_;
  set EconNew;
  by BRAND Segment ;
  TAB = ',' ;
  FN = "C:\SAMD2\CHAPTER2\ASSIGNMENTS\DatKgrp.CSV";
  file PLOTFILE filevar=FN;
  if ( FIRST.Brand ) then
  do;
    put "SEGMENT" TAB "BRAND" TAB "PERIOD" TAB "INTENT" TAB "SHARE" ;
  end;
    put  SEGMENT  TAB  BRAND  TAB   PERIOD TAB  Intent  TAB  Share ;
run;
```

Fig. 2.10 (continued)

Bibliography

Basic Technical Readings

Tatsuoka, Maurice M. (1971), "Multivariate Analysis: Techniques for Educational and Psychological Research," New York, NY: John Wiley & Sons, Inc.

Application Readings

Cool, Karel and Ingemar Dierickx (1993), "Rivalry, Strategic Groups and Firm Profitability," *Strategic Management Journal*, 14, 47–59.
Long, Rebecca G., William P. Bowers, Tim Barnett, et al. (1998), "Research Productivity of Graduates in Management: Effects of Academic Origin and Academic Affiliation," *Academy of Management Journal*, 41(6), 704–771.

Springer