

Chapter 2

Multivariate Normal Distribution

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

2.1 Univariate Normal Distribution

To review, in the case of a single random variable, the probability distribution or the density function of that variable x is represented by Eq. (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

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

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

Fig. 2.1 The bivariate normal distribution

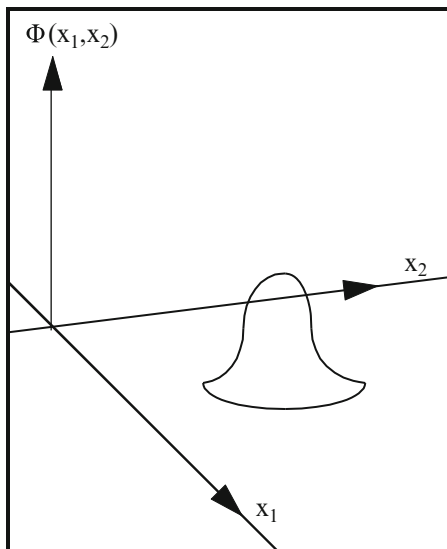
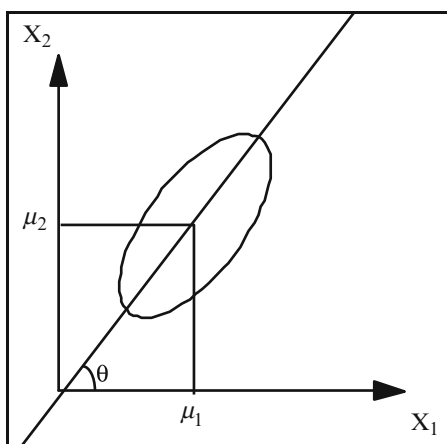


Fig. 2.2 The locus of points of the bivariate normal distribution at a given density level



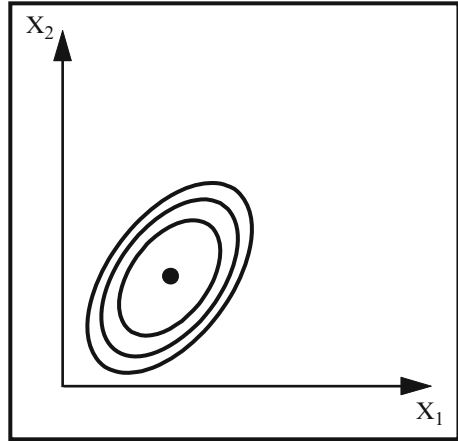
The *isodensity contour* is defined as the set of points for which the values of x_1 and x_2 give the same value for the density function Φ . This contour is given by Eq. (2.3) for a fixed value of C , which defines a constant probability:

$$\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} = C \tag{2.3}$$

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 . 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

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

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

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

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,

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

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

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

Now, more generally with p random variables (x_1, x_2, \dots, 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

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

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

$$x \sim N(\mu, \sigma^2) \quad (2.11)$$

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

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

2.4.1.2 Multivariate Distribution

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

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

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 $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}/n$:

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

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., $\mu_o = 150$) with the following information:

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

Then, the z score can be computed:

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

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

$$\left[154 - 1.96 \times \frac{16}{8}, 154 + 1.96 \times \frac{16}{8} \right] = [150.08, 157.92]$$

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

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}} \quad (2.15)$$

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:

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

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

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

$$H: \quad \boldsymbol{\mu}_o = \begin{bmatrix} 20 \\ 15 \end{bmatrix}$$

At the alpha level of 0.05, the value of the density function can be written as in Eq. (2.16), which follows a chi-square distribution at the specified significance level α :

$$n(\boldsymbol{\mu}_o - \bar{\mathbf{x}})' \Sigma^{-1} (\boldsymbol{\mu}_o - \bar{\mathbf{x}}) \sim \chi_p^2(\alpha) \quad (2.16)$$

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 2 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 $\boldsymbol{\mu} = \begin{bmatrix} 20 \\ 15 \end{bmatrix}$ is rejected.

2.4.2.3 Multivariate Test with Unknown $\boldsymbol{\Sigma}$

Just as in the univariate case, $\boldsymbol{\Sigma}$ is replaced with the sample value $\mathbf{S}/(n-1)$, where \mathbf{S} is the sums-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:

$$\text{Hotelling : } T^2 = n(n-1)(\bar{\mathbf{x}} - \boldsymbol{\mu}_o)' \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_o) \quad (2.17)$$

where if

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

then

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

Hotelling showed that

$$\frac{n-p}{(n-1)p} T^2 \sim F_{n-p}^p \quad (2.18)$$

Replacing T^2 by its expression given in Eq. (2.17) leads to

$$\frac{n(n-p)}{p} (\bar{\mathbf{x}} - \boldsymbol{\mu}_o)' \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_o) \sim F_{n-p}^p \quad (2.19)$$

Consequently, the test is performed by computing the expression in Eq. (2.19) and by 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

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

where

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

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:

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

The computations for testing the significance of the differences are similar to those in Sect. 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 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 ,

$$\mathbf{X}^{d(k)} = \begin{bmatrix} x_{11}^{(k)} - \bar{x}_1^{(k)} & x_{21}^{(k)} - \bar{x}_2^{(k)} & \dots \\ x_{12}^{(k)} - \bar{x}_1^{(k)} & x_{22}^{(k)} - \bar{x}_2^{(k)} & \dots \\ \vdots & \vdots & \ddots \\ x_{1n_k}^{(k)} - \bar{x}_1^{(k)} & x_{2n_k}^{(k)} - \bar{x}_2^{(k)} & \dots \end{bmatrix} \quad (2.23)$$

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

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

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

$$\mathbf{W} = \sum_{k=1}^K \mathbf{S}_k \quad (2.25)$$

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

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

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

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

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

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:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} \quad (2.29)$$

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

\mathbf{W} is defined as in Eq. (2.25). The total SSCP matrix is the sums of squares and 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

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

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

We create a new data matrix that comprises the centered data for each of the groups, stacked one upon the other:

$$\mathbf{X}^{d^*}_{n \times p} = \begin{bmatrix} \mathbf{X}^{d^*(1)} \\ \mathbf{X}^{d^*(2)} \\ \vdots \\ \mathbf{X}^{d^*(K)} \end{bmatrix} \quad (2.31)$$

The total SSCP matrix \mathbf{T} is then defined as

$$\mathbf{T}_{p \times p} = \mathbf{X}^{d^*t}_{p \times n} \mathbf{X}^{d^*}_{n \times p} \quad (2.32)$$

Intuitively, if we reduce the space to a single variate so that we are only dealing with variances and no covariances, Wilk's lambda (Λ) is the ratio of the pooled within-groups 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-groups variance. Consequently, the ratio Λ 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 Eq. (2.29), it follows that

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

and consequently,

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

Also, it follows that

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

When Wilk's Λ approaches 1, we showed that it means that the difference in means is negligible. This is the case when $\text{Ln } \Lambda$ approaches 0. However, when Λ approaches 0, it means that the difference is large. Therefore, a large value of $-\text{Ln}\Lambda$ is an indication of the significance of the difference between the means.

Based on Wilk's Λ , 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:

$$V = -[N - 1 - (p + K)/2]\text{Ln}\Lambda \sim \chi_{p(K-1)}^2 \quad (2.36)$$

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

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

$$R = \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{wt - p(K - 1)/2 + 1}{p(K - 1)} \approx F_{\nu_1=p(K-1), \nu_2=wt-p(K-1)/2+1} \quad (2.37)$$

where

$$w = N - 1 - (p + K)/2$$

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

The parameter t is set to 1 if either the numerator or the denominator of this last expression equals 0. The F statistic is exact when there are only one or two variables (p) or when the number of groups (K) equals 2 or 3.

A significant chi-square for Bartlett's test or a significant F test for Rao's test indicates significant differences in the group means.

2.5 Examples

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 whether or not the market share, distribution coverage, and prices are similar or

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:\SAMD\Chapter2\Examples\Mkt_Data.csv"
DIM = ' ', 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)

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 in Table 2.1.

The SAS file showing the SAS code needed to compute the necessary statistics is shown in Fig. 2.4. The first lines correspond to the basic SAS commands to read the

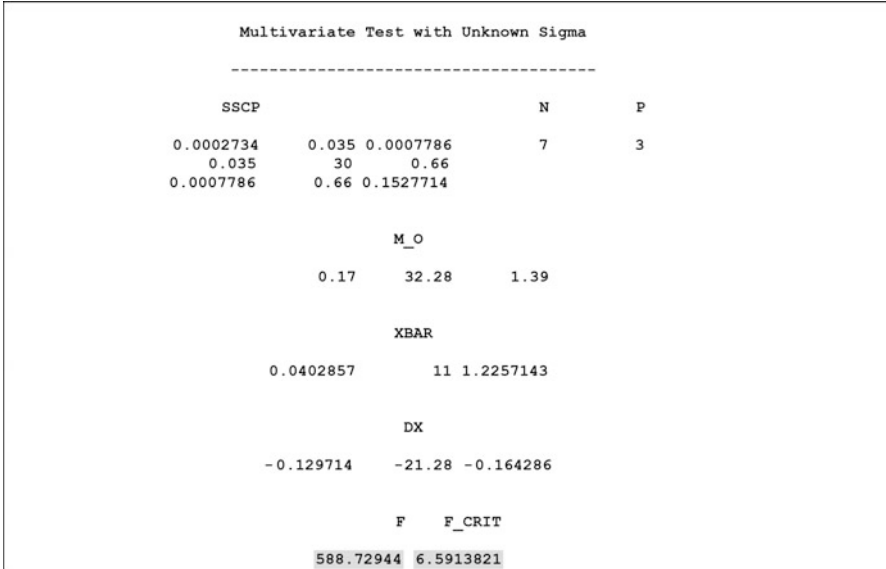


Fig. 2.5 SAS output of analysis defined in Fig. 2.4 (examp2-1.lst)

data from the file. Here, the data file was saved as a text file from Microsoft Excel. Consequently, the values in the file corresponding to different data points are separated by commas. This is indicated as the delimiter (“dlim”). Also, the data (first observation) start on line 2 because the first line is used for the names of the variables (as illustrated in Table 2.1). The variable PERIOD is dropped so that only the three variables needed for the analysis are kept in the SAS working data set. The IML procedure 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 commands, especially the file name and path and the variables. Within the IML subroutine, only two items would need to be changed: (1) the variables used for the analysis and (2) the values for the null hypothesis (m_o). The results are printed in the output file shown in Fig. 2.5.

The critical F statistic with 3 and 4 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 the United Kingdom) for seven periods, as shown in Table 2.2. The question is whether or not the mean vectors are the same for the three countries.

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

We first present an analysis that shows the matrix computations following precisely the equations presented in Sect. 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 reader who wishes to skip the detailed calculations can go directly to the SAS GLM procedure that is illustrated in Fig. 2.8.

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

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

These results indicate that the Bartlett's V statistic of 82.54 is larger than the critical chi-square with 6 degrees of freedom at the 0.05 confidence level ($\chi^2_{(df=6, \alpha=0.05)} = 12.59$). Consequently, the hypothesis that the mean vectors are the same is rejected. The same conclusion can be derived from 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 ($F^{\nu_1=6}_{\nu_2=32}(\alpha=0.05) = 2.399$).

The first lines of SAS commands 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 because the procedure automatically performs the MANOVA tests. For that analysis, the general procedure of the general linear model is called with the command "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

```

***** Examp2-2.sas ***** */
OPTIONS LS=80;
DATA work;
INFILE
"C:\SAMD\CHAPTER2\EXAMPLES\Mkt_Dt_K.csv"
dlim = ',' 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[+,1]/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)

in the variable. On the left side of the equal sign, the model statement shows the list of the variates for which the means will be compared. On the right side is the group variable. The GLM procedure is in fact a regression where the dependent variable is regressed on the dummy variables that are automatically created by SAS (different dummy variables are created for each of the values of the grouping variable).

```

-----
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      DEFW
      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:\SAMD\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 /noui;
manova h = CNTRY/ printe;
run;
quit;

```

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

The optional parameter “noui” 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 is necessary to indicate 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 Λ has the same value of 0.007787. Several other tests are provided, and they

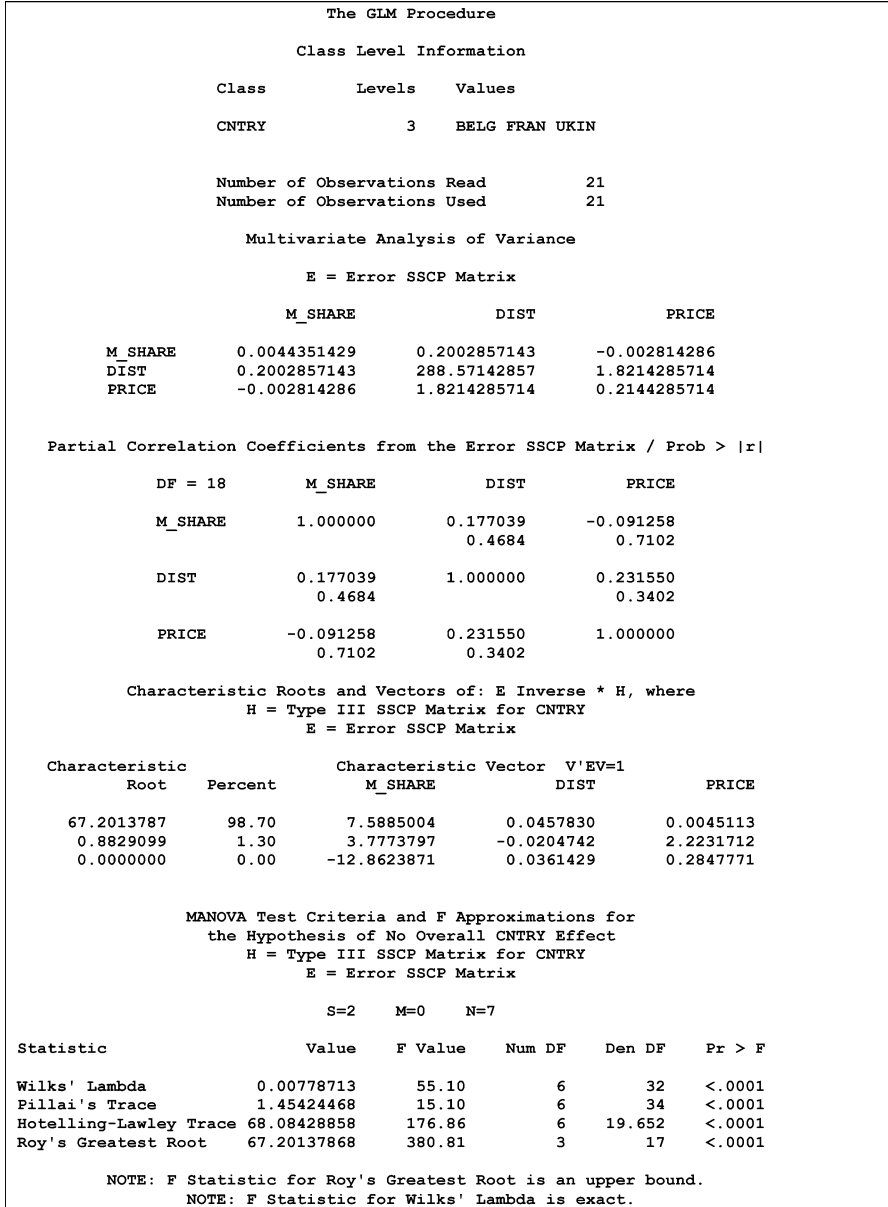


Fig. 2.9 SAS output for MANOVA test of mean differences across K groups (examp2-3.lst)

all lead to the same conclusion that the differences in means are significant. In addition to the expression of Wilk's Λ as a function of the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$, three other measures are provided in the SAS output.

Pillai's trace is defined as
$$\sum_{i=1}^K \frac{\lambda_i}{1 + \lambda_i}.$$

```

insheet using
"/users/gatignon/documents/WORK_STATA/SAMD/Chapter2_MANOVA/Mkt_Dt_K.csv", clear
* manova test
manova m_share dist price = centryno
mat list e(E)
mat list e(H_m)
mat list e(eigenvals_m)
mat list e(aux_m)

```

Fig. 2.10 STATA input for MANOVA test of mean differences across K groups (examp2-3.do)

Hotelling–Lawley trace is simply the sum of the eigenvalues: $\sum_{i=1}^K \lambda_i$.

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.

Similar output is provided by STATA. Figure 2.10 shows the input for requesting MANOVA analysis in STATA.

Figure 2.11 presents the results of the analysis. It includes the within- and the between-SSCP matrices. The command “mat list e(E)” is used to print the within-SSCP matrix and “mat list e(H_m)” the between-SSCP matrix. The largest root is read from the eigenvector computed by “e(eigenvals_m).” Finally, the command “mat list e(aux_m)” lists the parameters m , s , and n that are used for the F values corresponding to the various statistics shown in the output. These parameters are defined as follows:

$$s = \min(K - 1, p) \quad (2.38)$$

$$m = (|K - 1 - p| - 1)/2 \quad (2.39)$$

$$n = (N - K - p - 1)/2 \quad (2.40)$$

where

N = total number of observations across groups;

K = number of groups;

p = number of variables.

For example, an approximate F statistic for Pillai’s trace V with $s(2m + s + 1)$ and $s(2n + s + 1)$ degrees of freedom is

$$F = \frac{(2n + s + 1)V}{(2m + s + 1)(s - V)} \quad (2.41)$$

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:

```

. insheet using
"/users/gatignon/documents/WORK_STATA/SAMD/Chapter2_MANOVA/Mkt_Dt_K.csv", clear
(6 vars, 21 obs)

. * manova test
. manova m_share dist price = centryno

                Number of obs =      21

                W = Wilks' lambda      L = Lawley-Hotelling trace
                P = Pillai's trace      R = Roy's largest root

-----+-----
Source | Statistic  df  F(df1,  df2) =  F  Prob>F
-----+-----
centryno | W   0.0078    2    6.0   32.0   55.10  0.0000 e
          | P   1.4542    6.0   34.0   15.10  0.0000 a
          | L  68.0843    6.0   30.0   170.21 0.0000 a
          | R  67.2014    3.0   17.0   380.81 0.0000 u
-----+-----
Residual |                18
-----+-----
Total    |                20
-----+-----
e = exact, a = approximate, u = upper bound on F

. mat list e(E)

symmetric e(E)[3,3]
      m_share      dist      price
m_share .00443514
dist .20028564  288.57143
price -.00281429  1.8214294  .21442857

. mat list e(H_m)

symmetric e(H_m)[3,3]
      m_share      dist      price
m_share .12321343
dist 42.401429  14601.238
price .18368288  61.987627  .42906667

. mat list e(eigvals_m)

e(eigvals_m)[1,2]
      c1      c2
r1  67.201384  .88290968

. mat list e(aux_m)

e(aux_m)[3,1]
      value
s      2
m      0
n      7

```

Fig. 2.11 STATA output for MANOVA test of mean differences across K groups (examp2-3.log)

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 2.12 shows how to read the data within an SAS file and how to create new files with a subset of the data saved in a format that can be read easily using the examples provided throughout this chapter. Using the model described in the examples above, adapt these examples to the database to perform tests of differences across groups.

The commands to merge the INDUP and PANEL data sets in STATA are shown in Fig. 2.13.

```

/*****
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:\SAMD\Chapter2\Assignments\panel.csv' firstobs=2 dlm = ',' ;
input period segment segsize ideall-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:\SAMD\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:\SAMD\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 ;
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:\SAMD\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. 2.12 Example of SAS file for reading data sets INDUP and PANEL and creating new data files (assign2.sas)

```
insheet using "/users/fblgatignon/Documents/WORK_STATA/SAMD/panel.csv", clear
merge m:m period brand using "/users/fblgatignon/Documents/WORK_STATA/SAMD/indup.dta"
keep if segment ==5
drop if period ==0
regress awareness adshare
manova dolshare adshare relprice = firm
```

Fig. 2.13 Example of STATA file for reading and merging data sets INDUP and PANEL (MergeIndup_Panel_Mac.do)

Bibliography

Basic Technical Readings

Tatsuoka, M. M. (1971). *Multivariate analysis: techniques for educational and psychological research*. New York, NY: Wiley.

Application Readings

Cool, K., & Dierickx, I. (1993). Rivalry, strategic groups and firm profitability. *Strategic Management Journal*, 14, 47–59.

Kilduff, M., Angelmar, R., & Mehra, A. (2000). Top management-team diversity and firm performance: Examining the role of cognitions. *Organization Science*, 11(1), 21–34.

Long, R. G., Bowers, W. P., Barnett, T., et al. (1998). Research productivity of graduates in management: Effects of academic origin and academic affiliation. *Academy of Management Journal*, 41(6), 704–771.



<http://www.springer.com/978-1-4614-8593-3>

Statistical Analysis of Management Data

GATIGNON, H.

2014, XV, 563 p. 263 illus., 34 illus. in color., Hardcover

ISBN: 978-1-4614-8593-3