# Chapter-VIII

## INTRODUCTION TO MULTIVARIATE ANALYSIS

Multivariate statistical analysis is appropriate whenever several responses are measured on each object or experimental unit. Univariate analysis applied separately to each response leads to incorrect conclusions, since responses measured on the same object are generally correlated. Multivariate analysis can be simply defined as the application of statistical methods that deal with reasonably large number of characteristics or variables recorded on each object in one or more samples simultaneously. It provides statistical tools for the study of joint relationships of variables in data that contains intercorrelations. In other words, multivariate analysis differs from univariate and bivariate analysis in that it directs attention away from the analysis of the mean and variance of a single variable or from the pairwise relationship between two variables, to the analysis of the co-variances or correlations which reflect the extent of relationship among three or more variables. For example, a biometrician concerned with developing a taxonomy for classifying species of fowl on the basis of anatomical measurements may collect information on skull length, skull width, humerus length and tibia length.

**Remarks:**

1. The term objects in multivariate analysis refer to things, persons, individuals, events or in general entities on which the measurements are recorded. And the measurements relate to characteristics or attributes of the objects that are being recorded and in general are called variables.
2. Multivariate analysis investigates the dependency not only amongst the variables but also among the individuals on which observations are made.

### 8.1 Data Cube and Data Matrices:

In multivariate analysis, a researcher generally deals with data collected from ónô individuals, on ópô characters recorded over ólô locations or periods or groups. Thus, the basic input can be considered in terms of a data cube denoted by $x_{ijk}$, where

i = 1, 2, é .., n refers to objects

j = 1, 2, é .., p refers to characteristics/variables or attributes

k =1, 2, é .., L refers to location/periods

The data cube can be given a two dimensional representation by writing matrices within matrices as follows:

| | | **Characteristics** | | | |
|---|---|---|---|---|---|
| **Location/Period** | **Objects** | $X_1$ | $X_2$ | ….. | $X_p$ |
| I | $O_1$ | $x_{111}$ | $x_{121}$ | ….. | $x_1p_1$ |
| | $O_2$ | $x_{211}$ | $x_{321}$ | ….. | $x_2p_1$ |
| | : : | | | | |
| | $O_n$ | $x_{n11}$ | $x_{n21}$ | ….. | $x_np_1$ |
| II | $O_1$ | $x_{112}$ | $x_{122}$ | ….. | $x_1p_2$ |
| | $O_2$ | $x_{212}$ | $x_{222}$ | ….. | $x_2p_2$ |
| | : | | | | |
| | $O_n$ | $x_{n12}$ | $x_{n22}$ | ….. | $x_np_2$ |
| | ….. | ….. | ….. | ….. | ….. |
| $L^{th}$ | $O_1$ | $x_{11L}$ | $x_{12L}$ | ….. | $x_1p_L$ |
| | $O_2$ | $x_{21L}$ | $x_{22L}$ | ….. | $x_2p_L$ |
| | : | : | : | | : |
| | $O_n$ | $x_{n1L}$ | $x_{n2L}$ | ….. | $x_{npL}$ |

In many applications, the analyst considers only two coordinates of the data cube, so that the basic input becomes a data matrix or rectangular array of numerical entities. This may result from collecting information only on one occasion/location or groups. In such situations, the data matrix has ónô rows and ópô columns and can be described in terms of the elements $x_{ij}$, where i refers to objects and j refers to attributes.

A data matrix as ónô individuals recorded for ópô characters/variable can be described as:

| **Objects** | $X_1$ | $X_2$ | … | $X_p$ |
|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{12}$ | é | $x_1p$ |
| 2 | $x_{21}$ | $x_{22}$ | é | $x_2p$ |
| : : | : : | : : | : : | : : |
| n | $x_{n1}$ | $x_{n2}$ | é | $x_{np}$ |

or simply

$$x = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

## 8.2 Descriptive Measures in Multivariate Analysis:

Let X be a random vector of p-components and denoted by:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \quad \text{or} \quad X' = [x_1, x_2, ....., x_p]$$

The mean vector ($\mu$) covariance matrix ($\mathbf{\Sigma}$) and correlation matrix ($\boldsymbol{\rho}$) are given by:

$$\boldsymbol{\mu} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \qquad \sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

$$\text{and } \boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

where $\mu_i = E(X_i)$, $\mathring{u}_i^2 = V(X_i)$, $\mathring{u}_{ij} = Cov(X_i, X_j)$ and $\rho_{ij} = corr(X_i, X_j) = \dfrac{Cov(x_i, x_j)}{\sqrt{v(x_i)\, v(x_j)}}$

The mean vector ($\boldsymbol{\mu}$), covariance matrix ($\mathbf{\Sigma}$) and the correlation matrix ($\boldsymbol{\rho}$) given above respectively represent the measures of central tendency, dispersion and linear association for the p-dimensional multivariate population. The sample estimates for these measures may be obtained as follows:

Let **x** denote an n x p data matrix, where n is the number of observations and p is the number of variables. Then sample mean vector denoted by $\bar{\mathbf{x}}$ is given by:

$$\bar{\mathbf{x}}' = \frac{1}{n} 1' X = \frac{1}{n}(1,1,.....,1)\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$= \frac{1}{n}\left[\sum x_{i1} \quad \sum x_{i2} \quad ..... \quad \sum x_{ip}\right]$$

$$= \left[\bar{x}_1 \quad \bar{x}_2 \quad ..... \quad \bar{x}_p\right]$$

The sample covariance matrix: $\mathbf{S} = \frac{1}{n-1}\mathbf{X_d'X_d}$ where $\mathbf{X_d} = \mathbf{X - 1}\,\bar{X}'$ is the matrix of mean corrected scores and the matrix $x_d' x_d$ is often called the corrected sum of squares and product matrix.

The sample correlation matrix is usually denoted by $\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$ can be obtained by $\mathbf{R = D'SD}$ where **D** denote the diagonal matrix whose entries along the main diagonal are the reciprocals of the standard deviation of the variables in x, i.e.

$$\mathbf{D} = \begin{bmatrix} \frac{1}{s_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{s_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{s_p} \end{bmatrix} \text{ where } s_j^{\,2} = \frac{1}{n-1}\sum (x_{ij} - \bar{x}_j)^2$$

**Example-1:** Suppose the data matrix $\mathbf{x} = \begin{bmatrix} 2 & 3 & 3 \\ 3 & 2 & 6 \\ 4 & 6 & 5 \\ 3 & 5 & 6 \end{bmatrix}$ be a multivariate sample of size n=4 from a trivariate population.

Then sample mean vector:

$$\bar{\mathbf{x}}' = \frac{1}{4}[1,1,1,1]\begin{bmatrix} 2 & 3 & 3 \\ 3 & 2 & 6 \\ 4 & 6 & 5 \\ 3 & 5 & 6 \end{bmatrix} = \frac{1}{4}[12 \;\; 16 \;\; 20] = [3 \;\; 4 \;\; 5]$$

Now the mean corrected score matrix $\mathbf{x}_d$ is computed as:

$$\mathbf{X}_d = \mathbf{X} - \mathbf{1}\bar{\mathbf{X}}' = \begin{bmatrix} 2 & 3 & 3 \\ 3 & 2 & 6 \\ 4 & 6 & 5 \\ 3 & 5 & 6 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [3 \;\; 4 \;\; 5]$$

$$= \begin{bmatrix} 2 & 3 & 3 \\ 3 & 2 & 6 \\ 4 & 6 & 5 \\ 3 & 5 & 6 \end{bmatrix} - \begin{bmatrix} 3 & 4 & 5 \\ 3 & 4 & 5 \\ 3 & 4 & 5 \\ 3 & 4 & 5 \end{bmatrix} = \begin{bmatrix} -1 & -1 & -2 \\ 0 & -2 & 1 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

$$\therefore \mathbf{S} = \frac{1}{n-1}\mathbf{X}_d'\mathbf{X}_d = \frac{1}{4-1}\begin{bmatrix} -1 & 0 & 1 & 0 \\ -1 & -2 & 2 & 1 \\ -2 & 1 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & -1 & -2 \\ 0 & -2 & 1 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

$$= \frac{1}{3}\begin{bmatrix} 2 & 3 & 2 \\ 3 & 10 & 1 \\ 2 & 1 & 6 \end{bmatrix} = \begin{bmatrix} 2/3 & 1 & 2/3 \\ 1 & 10/3 & 1/3 \\ 2/3 & 1/3 & 2 \end{bmatrix} = \begin{bmatrix} 0.67 & 1 & 0.67 \\ 1 & 3.33 & 0.33 \\ 0.67 & 0.33 & 2 \end{bmatrix}$$

$$\text{Now } \mathbf{D} = \begin{bmatrix} 1/s_1 & 0 & 0 \\ 0 & 1/s_2 & 0 \\ 0 & 0 & 1/s_3 \end{bmatrix} = \begin{bmatrix} \sqrt{3/2} & 0 & 0 \\ 0 & \sqrt{3/10} & 0 \\ 0 & 0 & \sqrt{1/2} \end{bmatrix}$$

$$\therefore \mathbf{R} = \mathbf{D}'\mathbf{S}\mathbf{D} = \begin{bmatrix} \sqrt{3/2} & 0 & 0 \\ 0 & \sqrt{3/10} & 0 \\ 0 & 0 & \sqrt{1/2} \end{bmatrix}\begin{bmatrix} 2/3 & 1 & 2/3 \\ 1 & 10/3 & 1/3 \\ 2/3 & 1/3 & 2 \end{bmatrix}\begin{bmatrix} \sqrt{3/2} & 0 & 0 \\ 0 & \sqrt{3/10} & 0 \\ 0 & 0 & \sqrt{1/2} \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{2/3} & \sqrt{3/2} & \sqrt{2/3} \\ \sqrt{3/10} & \sqrt{10/3} & \frac{1}{\sqrt{30}} \\ \sqrt{2/3} & \frac{1}{\sqrt{18}} & \sqrt{2} \end{bmatrix}\begin{bmatrix} \sqrt{3/2} & 0 & 0 \\ 0 & \sqrt{3/10} & 0 \\ 0 & 0 & \sqrt{1/2} \end{bmatrix} = \begin{bmatrix} 1 & 3/\sqrt{20} & 1/\sqrt{3} \\ 3/\sqrt{20} & 1 & 1/\sqrt{60} \\ 1/\sqrt{3} & 1/\sqrt{60} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.67 & 0.58 \\ 0.67 & 1 & 0.13 \\ 0.58 & 0.13 & 1 \end{bmatrix}$$

## Important Multivariate Methods:

Multivariate data recorded on a large number of interrelated variables is often difficult to interpret. Therefore, there is a need to condense and sum up the essential features of the data through dimension reduction or some appropriate summary statistics for better interpretation. As a broad classification, the multivariate techniques may be classified as dependence methods and interdependence methods.

The methods in which one or more variables are dependent and others are independent are called dependant techniques. Multivariate regression, multivariate analysis of variance, discriminant analysis and canonical correlation analysis are the notable dependence techniques.

If interest centres on the mutual association across all the variables with no distinction made among the variable types, then such techniques are called interdependence techniques. Principal component analysis, factor analysis, cluster analysis and multi-dimensional scaling are the important interdependence techniques.

**Multivariate Regression:** It is concerned with the study of the dependence of one or more variables on a set of other variables called independent variables with the objective to estimate or predict the mean values of the dependent variables on the basis of the known values of the independent variables. If there is only one dependent variable and many independent variables, then it is known as multiple regression.

**Multivariate Analysis of Variance:** It is simply a generalization of univariate analysis of variance, where the primary objective is on testing for significant differences on a set of variables due to changes in one or more of the controlled (experimental) variables.

**Discriminant Analysis:** It is used to find linear combinations of the variables that separate the groups. Given a vector of p observed scores, known to belong to one of two or more groups, the basic problem is to find some function of the p scores (i.e. a linear combination) which can accurately assign individual with a given score into one of the groups.

**Principal Component Analysis:** It is a dimension reduction technique where the primary goal is to construct orthogonal linear combinations of the original variables that account for as much of the total variation as possible. The successive linear combinations

are extracted in such a way that they are uncorrelated with each other and account for successively smaller amounts of total variation.

**Cluster Analysis:** The purpose of cluster analysis is to reduce a large data set to meaningful subgroups of individuals or objects. The division is accomplished on the basis of similarity of the objects across a set of specified characteristics. The individuals of a particular subgroup or cluster are, in some sense, more similar to each other than to elements belonging to other groups.

**Canonical Correlation Analysis:** The most flexible of the multivariate technique, canonical correlation simultaneously correlates several explanatory variables and several dependent variables. In usual sense, it determines the linear association between a set of dependent variables and a set of explanatory variables. In canonical analysis, we find two linear combinations, one for the predictor set of variables and one for the set of explanatory variables, such that their product moment correlation is maximum.

**8.4 Testing Significance of Mean Vector (One Sample Case):** This is useful for multivariate populations where it is required to test whether the population mean vector ($\mu$) is equal to a specified mean vector ($\mu_0$).

Let $\mathbf{x}_1, \mathbf{x}_2$, é .., $\mathbf{x}_n$ be a random sample of size n from a p-dimensional multivariate population having mean vector $\mu$ and covariance matrix $\mathbf{\Sigma}$,

$H_0$: $\mu = \mu_0$

$H_1$: $\mu$ Í $\mu_0$

If covariance matrix Ɇ is known or sample is large then a $\chi^2$ test is used to test the above hypothesis. If $\overline{\mathbf{X}}$ denotes the sample mean vector then the statistic $\chi^2 = n\left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)' \sum^{-1} \left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)$ follows a chi-square distribution with p degrees of freedom and we reject $H_0$ if $\chi^2_{cal} > \chi^2(p)$ at desired Ŭ level of significance.

If the covariance matrix $\mathbf{\Sigma}$ is not known and sample size is small, then Hotelling $T^2$ (defined below) is used for testing $H_0$.

$T^2 = n\left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)' S^{-1} \left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)$, where $\mathbf{S}$ is the sample covariance matrix. The sampling distribution of Hotelling $T^2$ is given as:

$\frac{(n-p)}{(n-1)p}T^2 \approx F_{(p,\ n-p)}$ distribution or $T^2 \approx \frac{(n-1)p}{n-p}F_{(p,\ n-p)}$ distribution.

Therefore, we reject $H_0$ if $T^2_{cal} > \frac{(n-1)p}{n-p}F_{(p,\ n-p)}$ at Ŭlevel of significance.

**8.5 Testing Equality of Two Mean Vectors (Two sample case):** Consider two independent samples of sizes $n_1$ and $n_2$ from two p-variate normal populations with mean vectors $\mu_1$ and $\mu_2$ and having same but known covariance matrix $\mathbf{\Sigma}$.

$H_0$: $\mu_1 = \mu_2$

$H_1$: $\mu_1$ Í $\mu_2$

Test statistic is:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right)' \mathbf{S}_p^{-1}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right) \text{follows}$$

$$\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F(p, n_1 + n_2 - p - 1) \text{ distribution.}$$

where $\overline{\mathbf{X}}_1$ *and* $\overline{\mathbf{X}}_2$ are the sample mean vectors and $S_p$ is the pooled covariance matrix defined by $\mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$, $\mathbf{S}_1$ and $\mathbf{S}_2$ being the individual sample covariance matrices.

**Example-2:** Mean vector and covariance matrix for a sample of size 20 from a trivariate population are found to be:

$$\overline{\mathbf{X}} = \begin{bmatrix} 4.64 \\ 45.4 \\ 9.94 \end{bmatrix} \text{ and } \mathbf{S} = \begin{bmatrix} 2.88 & 10.01 & -1.81 \\ 10.01 & 199.79 & -5.64 \\ -1.81 & -5.64 & 3.63 \end{bmatrix}$$

We wish to test the hypothesis that the null hypothesis $H_0$ : $\mu = \mu_0$ where $\mu_0 = \begin{bmatrix} 4 \\ 50 \\ 10 \end{bmatrix}$ against $H_1$ : $\mu$ Í $\mu_0$.

$$T^2 = n(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}_0)$$

$$= 20\begin{bmatrix}4.64-4 & 45.4-50 & 9.97-10\end{bmatrix}\begin{bmatrix}2.88 & 10.01 & -1.81\\ 10.01 & 199.79 & -5.64\\ -1.81 & -5.64 & 3.63\end{bmatrix}^{-1}\begin{bmatrix}4.64-4\\ 45.4-50\\ 9.97-10\end{bmatrix} \simeq 9.7$$

For p = 3 and n = 20 and $\alpha = 0.05, \frac{(n-1)p}{n-p} F_{\alpha(p,n-p)} = 10.7$

Since $T^2_{cal}$ is less than 10.7, we do not reject $H_0$.

**Example-3:** Consider the two samples of size $n_1$ = 45 and $n_2$ = 55 with $\overline{\mathbf{X}}_1 = \begin{bmatrix}204.4\\ 556.6\end{bmatrix}$, $\overline{\mathrm{X}}_2 = \begin{bmatrix}130\\ 355\end{bmatrix}$

$\mathbf{S}_1 = \begin{bmatrix}13825.3 & 23823.4\\ 23823.4 & 73107.4\end{bmatrix}$ and $\mathbf{S}_2 = \begin{bmatrix}8632 & 19616.7\\ 19616.7 & 55964.5\end{bmatrix}$

Test $H_0$: $\mu_1 = \mu_2$ against $H_1$: $\mu_1$ Í $\mu_2$

After simplification, we get:

$$\mathbf{S}_p = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2} = \begin{bmatrix}10963.7 & 21505.5\\ 21505.3 & 63661.3\end{bmatrix}$$

and $T^2_{cal} = \frac{n_1 n_2}{n_1+n_2}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right)' \mathbf{S}_p^{-1}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right) = 16.2$

Comparing the calculated $T^2$ value with the critical value at Ŭ= 0.05 that is

$$\frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{(p,n_1+n_2-p-1)} = \frac{98 \text{ x } 2}{97} F_{0.05}(2,97) = 6.26$$

Since $T^2_{cal} > 6.26$, therefore, we reject $H_0$.

**Note:**

1. Mahalanobis $D^2$ defined below can also be used for testing the above hypothesis:

$$D^2 = \left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right)' \mathbf{S}_p^{-1}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right) = \frac{n_1 n_2}{n_1+n_2} T^2$$

2. If the two populations do not have same covariance matrices, then above test cannot be used. However, for large samples or when dispersion matrices are known, $\chi^2$ test can be used.