Chapter 1
Introduction to Multivariate Analysis

The objective of any statistical analysis is to find the information hidden inside the data. The data contains a lot of information, but the data itself cannot speak for itself. Various statistical tools are applied to the data to extract information.

A **random variable** is defined and associated with the characteristic of the data to be studied. For example, information on the heights of children is collected by defining a random variable "height" and collecting data on the heights of a certain number of children. Similarly, in another example, if the average performance of the car is to be studied in terms of mileage, then the random variable is defined as the "total distance travelled by the car per litre of fuel". A certain number of cars are selected, and data on the total distance travelled by each car per litre of fuel are recorded. Such an analysis, which uses only one random variable, is called **univariate** analysis. In a univariate analysis, it is assumed that only one variable affects the process.

When two random variables jointly affect the outcome, the analysis is called as a **bivariate** analysis. For example, a child's height and weight reflect their growth. To study the growth of children, data on children's height and weight are collected, and hence two random variables - "height of children" and "weight of children"- are defined. The data are paired, i.e., each pair consists of a child's height and weight. Such a data set will facilitate the study of the effects of height and weight on growth, as well as their joint effect. Thus, bivariate analysis facilitates the study of the effects of individual variables and their interactions or joint effects. For example, variances, covariances, and/or correlation coefficients facilitate the study of variation in bivariate data.

The concept of bivariate analysis can be extended and generalized to multivariate analysis. Similar to bivariate analysis, which considers two random variables, multivariate analysis considers more than two variables. For example, if a person's health is to be studied, the effects of several variables, such as age, height, weight, blood pressure, blood sugar, cholesterol level, etc., will jointly determine a person's health status. The setup of multivariate analysis involves more than two variables and analyzes the data obtained on

multiple variables simultaneously. This helps in understanding the effects of individual variables, as well as the relationships and structures in the data. It reveals the interaction patterns among variables, helps group observations, and reduces data complexity. For example, the behaviour of two or three variables simultaneously provides a realistic view of the data, underlying relationships, and patterns.

Multivariate analysis encompasses various statistical techniques that play an important role in Statistics and Data Science. **Dependence methods** in multivariate analysis examine cause-and-effect relationships and are useful when one or more variables depend on others. For example, the yield of a crop depends upon the quantity of fertilizer, irrigation levels, quantity of seeds, area under cultivation, etc. Such a technique helps in building predictive models in machine learning. **Interdependence methods** explore the structure of the dataset and help understand its patterns. For example, interdependence methods help identify sets of similar variables that can be grouped together.

Data Representation

Real-world datasets are mostly multivariate. Whatever is happening is the result of multiple inputs and influences. The data in the multivariate statistical analysis consist of sets of observations of measurements on a number of individuals or objects. The sample data, e.g., will be on the heights, weights and ages of randomly selected students from a school.

Data for a univariate analysis are collected on a single random variable, say X, and n observations are represented as $x_1, x_2, \ldots, x_n$. For example, if X is the height of persons, then $x_1 = 150$ cms. and $x_2 = 155$ cms. are two observations on the heights of two randomly selected persons. Similarly, the data in a bivariate analysis is collected on two random variables, say X_1 and X_2. For example, if X_1 and X_2 are the height and weight of a person, then $x_{11} = 150$ cms. and $x_{12} = 50$ Kg. are the height and weight of a randomly selected person and is represented as $\underset{\sim}{x}_1 = \Big(x_{11}\ ,\ x_{12}\Big)' = \Big(150\ ,\ 50\Big)'$. Similarly, the second observation on the height and weight of another randomly selected

person is represented as $\underset{\sim}{x}_2 = \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix}$ and so on. The corresponding random vector is represented as $\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ which is a 2×1 vector. On the similar lines, if there are p random variables $X_1, X_2, \ldots, X_p$, and if they are to be considered simultaneously, then they are represented as a random vector denoted by $\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = (X_1, X_2, \ldots, X_p)'$, which is a p-component vector or a $p \times 1$ column vector. The sample of n randomly selected observation on $\underset{\sim}{X}$ is represented as $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$ where each of the $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$ is a $p \times 1$ column vector. One can think of the entire vector as a randomly collected observation from a multivariate population. Such an outcome can be represented using a notation with two subscripts- one subscript for the variable and another subscript for the observation number. Let x_{ij} be the i^{th} observation on the j^{th} variable, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, p$. Consequently, the n measurements on p variables can be presented as

		Variables			
		X_1	X_2	$\ldots$	X_p
Observation number	1	x_{11}	x_{12}	$\ldots$	x_{1p}
Observation number	2	x_{21}	x_{22}	$\ldots$	x_{2p}
$\vdots$		$\vdots$	$\vdots$	$\ddots$	$\vdots$
Observation number	n	x_{n1}	x_{n2}	$\ldots$	x_{np}

and represented in an $n \times p$ matrix X as

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

The sample mean and sample variance for each of the p variables can be calculated from n observations. The sample covariances are calculated for each pair of random

variables. So there will be p sample means as

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad j = 1, 2, \ldots, p$$

and p sample variances as

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \quad j = 1, 2, \ldots, p.$$

The sample covariance of n observations for each pair of variables is

$$s_{jm} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m)$$

for $j \neq m = 1, 2, \ldots, p$.

It is to be noted that sometimes the sample variances and sample covariances are defined with a divisor $(n-1)$ as

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \quad \text{and} \quad \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m),$$

respectively, which are unbiased estimators for the respective population variances and covariances.

The sample correlation coefficient of n observations for each pair of variables is obtained as

$$r_{jm} = \frac{s_{jm}}{\sqrt{s_{jj}s_{mm}}}$$

for $j \neq m = 1, 2, \ldots, p$.

The notation of sample means, sample variances, sample covariances, and sample correlation coefficients can be extended to represent them in vector and matrix formats.

Sample Mean Vector

The sample mean vector is defined as a $p \times 1$ vector

$$\underset{\sim}{\bar{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)'$$

Sample Covariance Matrix

The sample covariance matrix of order $p \times p$ is defined as

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}$$

where the variances are mentioned on the diagonal elements, and covariances are mentioned on the off-diagonal elements of S.

Sample Correlation Matrix

The sample correlation matrix of order $p \times p$, denoted as as R, is defined as

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

where the correlation coefficients are defined on the off-diagonal elements. The diagonal elements will always be 1, since they represent each variable's correlation with itself.

Note that both S and R are symmetric matrices.

Example

Suppose the following observations are collected on the number of chocolates sold and their respective prices in four shops.

Thus, the observations for the first variable (x_1) are

$$x_{11} = 4, \quad x_{21} = 3, \quad x_{31} = 4, \quad x_{41} = 5$$

and the observations for the second variable (x_2) are

$$x_{12} = 48, \quad x_{22} = 58, \quad x_{32} = 42, \quad x_{42} = 52.$$

Shop Number	**Number of chocolates sold (Variable 1)**	**Cost of chocolate (in Rs.) (Variable 2)**
1	4	48
2	3	58
3	4	42
4	5	52

The sample observation vectors on the random vector $\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = (X_1 \quad X_2)'$ are

$$\underset{\sim}{x}_1 = \begin{pmatrix} 4 \\ 48 \end{pmatrix}, \quad \underset{\sim}{x}_2 = \begin{pmatrix} 3 \\ 58 \end{pmatrix}, \quad \underset{\sim}{x}_3 = \begin{pmatrix} 4 \\ 42 \end{pmatrix}, \quad \underset{\sim}{x}_4 = \begin{pmatrix} 5 \\ 52 \end{pmatrix}.$$

The sample means of observations on X_1 and X_2 are given as

$$\bar{x}_1 = \frac{1}{4}\sum_{i=1}^{4} x_{i1} = \frac{1}{4}(4+3+4+5) = \frac{16}{4} = 4$$
$$\bar{x}_2 = \frac{1}{4}\sum_{i=1}^{4} x_{i2} = \frac{1}{4}(48+58+42+52) = \frac{200}{4} = 50.$$

Thus, the sample mean vector is obtained as follows:

$$\underset{\sim}{\bar{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 50 \end{pmatrix}.$$

The sample variances and covariance are obtained as follows:

Variance of Variable 1 (s_{11})

$$\begin{aligned} s_{11} &= \frac{1}{4}\sum_{i=1}^{4}(x_{i1} - \bar{x}_1)^2 \\ &= \frac{1}{4}\left[(4-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2\right] \\ &= \frac{1}{4}[0+1+0+1] = \frac{2}{4} = 0.5. \end{aligned}$$

Variance of Variable 2 (s_{22})

$$\begin{aligned} s_{22} &= \frac{1}{4}\sum_{i=1}^{4}(x_{i2}-\bar{x}_2)^2 \\ &= \frac{1}{4}\left[(48-50)^2+(58-50)^2+(42-50)^2+(52-50)^2\right] \\ &= \frac{1}{4}[(-2)^2+(8)^2+(-8)^2+(2)^2] \\ &= \frac{1}{4}[4+64+64+4] = \frac{136}{4} = 34. \end{aligned}$$

Covariance between X_1 and X_2 ($s_{12} = s_{21}$)

$$\begin{aligned} s_{12} &= \frac{1}{4}\sum_{i=1}^{4}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) \\ &= \frac{1}{4}[(4-4)(48-50)+(3-4)(58-50)+(4-4)(42-50)+(5-4)(52-50)] \\ &= \frac{1}{4}[0+(-1)(8)+0+(1)(2)] \\ &= \frac{1}{4}[-8+2] \\ &= \frac{-6}{4} = -1.5. \end{aligned}$$

Sample Covariance Matrix of X_1 and X_2 (S)

The sample covariance matrix S of X_1 and X_2 is obtained as

$$S = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = \begin{pmatrix} 0.5 & -1.5 \\ -1.5 & 34 \end{pmatrix}.$$

Sample Correlation Coefficient (r_{12})

The sample correlation r_{12} is obtained as

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}s_{22}}} = \frac{-1.5}{\sqrt{34\times 0.5}} = \frac{-1.5}{\sqrt{17}} \approx -0.36.$$

Note that $r_{12} = r_{21}$.

Sample Correlation Matrix of X_1 and X_2 (R)

The sample correlation matrix of X_1 and X_2 is obtained as

$$R = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} = \begin{pmatrix} 1 & -0.36 \\ -0.36 & 1 \end{pmatrix}.$$

Joint Distributions

For two random variables X and Y, the cumulative distribution function (CDF) is defined as

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

for every pair of real numbers (x, y). When $F_{XY}(x, y)$ is absolutely continuous, i.e., the following partial derivative exists almost everywhere

$$\frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} = f_{XY}(x, y)$$

and

$$F_{XY}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v) du dv.$$

The nonnegative function $f_{XY}(x, y)$ is the probability density function (PDF) of X and Y.

For p random variables, $X_1, X_2 \ldots, X_p$, the CDF is defined as

$$F_{X_1, X_2, \ldots, X_p}(x_1, x_2, \ldots, x_p) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_p \leq x_p)$$

is defined for every set of real numbers $x_1, x_2 \ldots, x_p$. The PDF of $X_1, X_2, \ldots, X_p$, if $F_{X_1, X_2, \ldots, X_p}(x_1, x_2, \ldots, x_p)$ is absolutely continuous, is

$$\frac{\partial^p F_{X_1, X_2, \ldots, X_p}(x_1, x_2, \ldots, x_p)}{\partial x_1 \partial x_2, \ldots, \partial x_p} = f_{X_1, X_2, \ldots, X_p}(x_1, x_2, \ldots, x_p),$$

almost everywhere and

$$F_{X_1, X_2, \ldots, X_p}(x_1, x_2, \ldots, x_p) = \int_{-\infty}^{x_p} \int_{-\infty}^{x_{p-1}} \ldots \int_{-\infty}^{x_1} f(u_1, u_2, \ldots, u_p) du_1, du_2, \ldots, du_p.$$

The probability of falling in any (measurable) set $\mathbb{R}$ in the p-dimensional Euclidean space is

$$P[(X_1, X_2, \ldots, X_p) \in \mathbb{R}] = \int\int_{\mathbb{R}} \ldots \int f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) dx_1 dx_2 \ldots dx_p.$$

The joint moments of a subset of variates are defined as

$$\begin{aligned} & E\big(X_1^{m_1} X_2^{m_2} \ldots X_p^{m_p}\big) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_1^{m_1} x_2^{m_2} \ldots x_p^{m_p} \, f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) \, dx_1 dx_2 \ldots dx_p. \end{aligned}$$

For convenience, sometimes $f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)$ and $F_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)$ are written as $f(x_1, x_2, \ldots, x_p)$ and $F(x_1, x_2, \ldots, x_p)$, respectively.

Marginal Distributions

For a given joint CDF of random variables X and Y, the marginal CDF of X is

$$P(X \le x) = P(X \le x, \; Y \le \infty) = F_{X,Y}(x, \infty).$$

If the joint PDF exists, then

$$\begin{aligned} F_X(X) &= \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(u, v) \, dv \, du \\ &= \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f_{X,Y}(u, v) \, dv\right] du \\ &= \int_{-\infty}^{x} f_X(u) \, du. \end{aligned}$$

Similarly, the marginal CDF of Y is

$$G_Y(y) = \int_{-\infty}^{y} g_Y(v) \, dv.$$

Here, $f_{X,Y}(u, v)$ is the joint probability density function, $f_X(u)$ and $g_Y(v)$ are the marginal density functions of X and Y, respectively.

For a given joint CDF $F_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)$ of p random variables $X_1, X_2, \ldots, X_p$, the marginal distribution of $X_1, X_2, \ldots, X_r, \;\; (r < p)$, is given by

$$\begin{aligned} P(X_1 \le x_1, \ldots, X_r \le x_r) &= P(X_1 \le x_1, \ldots, X_r \le x_r, X_{r+1} \le \infty, \ldots, X_p \le \infty) \\ &= F_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_r, \infty, \ldots, \infty). \end{aligned}$$

Marginal density of a Subset of Random Variables

If the joint density function $f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)$ exists, then the marginal density of $X_1, X_2, \ldots, X_r$ is

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(u_1, u_2 \ldots, u_p)\, du_{r+1} du_{r+2} \ldots du_p.$$

Joint Moments of a Subset of Variables

The joint moments of a subset of variates are defined as

$$\begin{aligned}
&E(X_1^{m_1} X_2^{m_2} \cdots X_r^{m_r}) \\
&= E\big(X_1^{m_1} X_2^{m_2} \cdots X_r^{m_r} X_{r+1}^0 \ldots X_p^0\big) \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{m_1} x_2^{m_2} \cdots x_r^{m_r}\, f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)\, dx_1 dx_2 \cdots dx_p \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{m_1} x_2^{m_2} \cdots x_r^{m_r} \\
&\quad \times \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)\, dx_{r+1} \ldots dx_p\right] dx_1 dx_2 \cdots dx_r.
\end{aligned}$$

Statistical Independence of Random Variables

The random variables X and Y are said to be independent if

$$F_{XY}(x, y) = F_X(x) \cdot F_Y(y),$$

where $F_X(x)$ and $F_Y(y)$ are the marginal CDF of X and Y, respectively. This implies

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y).$$

Conversely, if $f_{XY} f(x, y) = f_X(x) \cdot g_Y(y)$ then $F_{XY}(x, y) = F_X(x) \cdot F_Y(y)$.

The p random variables $X_1, X_2, \ldots, X_p$ with CDF $F_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)$ are said to be mutually independent if

$$F_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdots F_{X_p}(x_p),$$

where $F_{X_i}(x_i)$ is the CDF of X_i, $i = 1, 2, \ldots, p$.

The set $X_1, X_2, \ldots, X_r$ is said to be independent of the set $X_{r+1}, X_{r+2}, \ldots, X_p$ if

$$F_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) = F_{X_1,X_2,\ldots,X_r}(x_1, x_2, \ldots, x_r) \times F_{X_{r+1},X_{r+2},\ldots,X_p}(x_{r+1}, x_{r+2}, \ldots, x_p).$$

Also, if $X_1, X_2, \ldots, X_p$ are mutually independent then

$$E(X_1^{m_1} X_2^{m_2} \cdots X_p^{m_p}) = E(X_1^{m_1})\, E(X_2^{m_2}) \cdots E(X_p^{m_p}).$$

Conditional Distributions

Let X and Y be two random variables. The conditional distribution probability that X falls in (x_1, x_2) given that Y falls in (y_1, y_2) is

$$P(x_1 \leq X \leq x_2 \mid y_1 \leq Y \leq y_2) = \frac{\displaystyle\int_{x_1}^{x_2} \int_{y_1}^{y_2} f(u, v)\, dv\, du}{\displaystyle\int_{y_1}^{y_2} g(v)\, dv}.$$

For p random variables $X_1, X_2, \ldots, X_p$ with CDF $F_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)$, the conditional density of $X_1, X_2, \ldots, X_r$ given $X_{r+1} = x_{r+1}, \ldots, X_p = x_p$ is

$$\frac{f(x_1, x_2, \ldots, x_p)}{\displaystyle\int_{-\infty}^{\infty}, \int_{-\infty}^{\infty}, \ldots, \int_{-\infty}^{\infty} f(u_1, u_2, \ldots, u_r, x_{r+1}, \ldots, x_p)\, du_1\, du_2 \cdots du_r}.$$

Note: Univariate Versus Multivariate In multivariate analysis, it is preferred to use vectors and matrices. Initially, it may not be comfortable for a reader to think directly in terms of vectors and matrices in comparison of scalars, but with practice, it is not difficult to achieve it over time. For example, $\sum_{i=1}^{n} x_i^2$ is a scalar, but it can be expressed in terms of vectors as $\underset{\sim}{x}'\underset{\sim}{x}$ where $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)'$ is an $n \times 1$ vector.

Similarly, $2x_1^2 + 2x_1x_2 + 3x_2^2$ is expressed as

$$(x_1\ x_2) \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \underset{\sim}{x}' A \underset{\sim}{x}$$

where $A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$.