Unit I

Introduction:-

Multivariate statistical analysis is concerned with data that consist of sets of measurements (variables or characters) on no. of individuals (or) objects. The sample data may be heights and weights of some individuals drawn randomly from a popln of school children in a given city.

In general, multivariate data arise whenever an investigator, seeking to understand a social (or) physical phenomenon, checks a number $p \geq 2$ of variables (or) characters (or) measurement to record. The values of these variables are all recorded for each distinct item (or) individual (or) experimental unit.

We will use the notation x_{jk} to indicate the particular value of the k th variable that is observed on the j th item or trial. That is,

x_{jk} = measurement of the k th variable on j th item.

Consequently, n objects on p variables can be displayed as follows.

	Variable 1	....	Variable k	...	Variable
Item 1	x_{11}	....	x_{1k}	....	x_{1p}
Item 2	x_{21}	....	x_{2k}	....	x_{2p}
⋮	⋮		⋮		⋮
Item j	x_{j1}	....	x_{jk}	....	x_{jp}
⋮	⋮		⋮		⋮
Item n	x_{n1}	....	x_{nk}	....	x_{np}

The measurement made on a single individual can be assembled in a row vector. we think of the entire vector as an observation from a multivariate population (or) distribution. when the individual is drawn randomly, we consider the vector as a random vector with a distribution or probability law describing that population.

The set of observations on all individuals in a sample constitutes a sample of vectors, and the vectors set side by side make up the matrix of observations. The data to be analysed then are thought of as displayed in a table (or) in several tables.

characteristics of a univariate distn of essential interest are the mean as a measure of location and the S.D as a measure of variability; similarly the mean and S.D of a univariate sample are important summary measures

In multivariate analysis, the means and variances of the separate measurements - for distn and for samples - have corresponding relevance. An essential aspect, however, of multivariate analysis is dependence b/w the different variables. The dependence b/w two variables may involve the covariance b/w them, that is, the average products of their deviations from their respective means. The covariance standardised by the corresponding S.D is the correlation coefficient, serves as a measure of degree of dependence.

A set of summary statistics is the mean vector (consisting of the univariate means) and the covariance matrix (consisting of univariate ~~means~~ variances and bivariate covariances) An alternative set of summary statistics with the same information is the mean vector, the set of S.D and the correlation matrix. Similar parameter quantities describe location variability and dependence in the popln or for a prob. distn The multivariate normal distn is completely described by its mean vector and covariance matrix. Covariance matrix constitute a sufficient set of statistics.

Application of multivariate Techniques:-

The published applications of multivariate methods have increased tremendeously in recent years. It is

1. Data reduction and simplification

(a) Using Data on several variables related to cancer patient responses to radiotherapy, a simple measure of patient response to radiotherapy was constructed

(b) Track records from many nations were used to develop an index of performance for both male and female athletes.

(c) Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images of a shoreline in two dimensional.

(d) Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants

(e) A matrix of tactic similarities was developed from aggregate data derived from professional mediators. From this matrix the no. of dimensions by which professional mediators judge the tactics they use in resolving dispute was determined.

2) Sorting and Grouping

(a) Data on several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing computer utilization.

(b) Measurement of several physiological variables were used to develop a screening procedure that discriminates alcoholics from non-alcoholics.

(c) Data related to responses to visual stimuli were used to develop a rule of separating people suffering from a multiple-sclerosis-caused visual pathology from those not suffering from the disease.

(d) The US internal revenue services needs data collected from tax returns to sort taxpayers into 2 groups those will be audited and those will not.

3) Investigation of the dependence among variables:-

(a) Data on several variables were used to identify that were responsible for client success in hiring external consultants.

(b) measurement of variables related to innovation, on the one hand and variables related to the business environment and business organisation, on the other hand, were used to discover why some firms are product innovators and some firm are not.

(c) measurement of pulp fiber characteristics and subsequent measurements of characteristics of the paper made from them are used to examine the relations b/w pulp fiber properties and the resulting paper properties. The goal is to determine those fibres that lead to higher quality paper.

(d) The association b/w measures of risk-taking propensity and measure of socio economic characteristic for top-level business executives were used to assess the relation b/w risk-taking behaviour and performance.

4) Prediction:-

(a) The association b/w test scores and high school performance variables and several college performance variables were used to develop predictors of success in college.

(b) measurements on several accounting and financial variables were used to develop a method for identifying potentially insolvent property-liability insurers.

5) Hypothesis Testing:-

(a) several pollution related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week or whether there is a noticeable difference b/w weekdays and weekends.

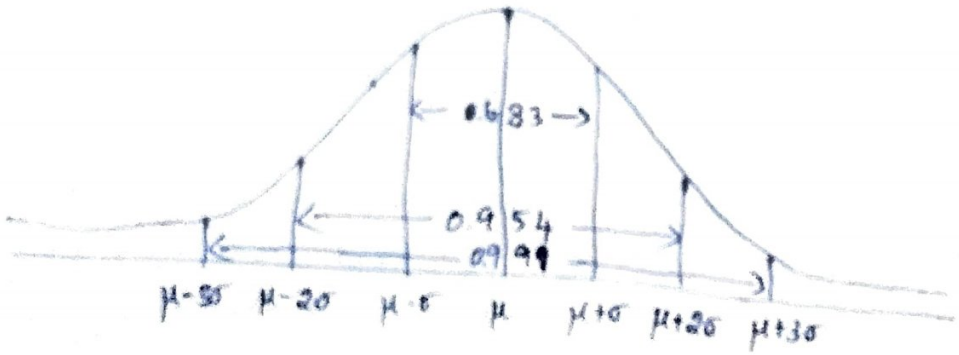
(b) data on several variable were used to determine whether different types of firms in newly industrialised countries exhibited different patterns of innovations

These above descriptions often glimps into use of multivariate methods in widely diverse fields.

The multivariate normal distribution and its properties

The multivariate normal density is a generalisation of the univariate normal density to $p \geq 2$ dimensions Recall that the univariate normal distribution, with mean μ and variance σ^2 has the prob distn function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ —(1)

$-\infty < x < \infty$



The above figure shows a normal density with mean μ and variance σ^2 and selected areas under the curve A plot of this function yields the familiar bell-shaped curve shown in the above figure Also shown in the figure are approximate areas under the curve within ± 1, ± 2 and ± 3 S.D of the mean These areas represents prob and thus

for the normal random variables,

$$P[\mu-\sigma \le x \le \mu+\sigma] = 0.68 \quad P[\mu-2\sigma \le x \le \mu+2\sigma] = 0.95$$

$$P[\mu-3\sigma \le x \le \mu+3\sigma] = 0.99$$

It is convinient to denote the normal density function with mean μ and variance σ^2 by $N(\mu, \sigma^2)$. Therefore $N(10,4)$ refers to function ① with $\mu=10$ and $\sigma=2$

This notation will be extended to the multivariate case

The term

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu) \quad \text{—②}$$

in the expondent of the univariate normal density function measures the square of the distance from x to μ in S.D units This can be generalised for $p \times 1$ vector x of observations on several variables as $(x-\mu)' \Sigma^{-1} (x-\mu)$ — ③

The $p \times 1$ vector μ represents the expected value of r.vector x and the $p \times p$ matrix Σ is the variance - covariance matrix of x. we shall assume that the symmetric matrix Σ is +ve definite. so the expression in ③ is the square of the generalised distance from x to μ

The multivariate normal density is obtained by replacing the univariate distance ② by the multivariate

generalised distance of ③ in the density function of ①

when this replacement is made, the univariate normalising constant $(2\pi)^{-1/2}(\sigma^2)^{-1/2}$ must be changed to a more general constant that makes the volume under the surface of the multivariate density function unity for any p This is necessary because, in multivariate case, probs are represented by volumes under the surface over regains defined by interval of the x_i values

It can be shown that the constant is $(2\pi)^{-p/2}|\Sigma|^{-1/2}$

and consequently, a p-dimensional normal density for the random vector $x' = [x_1, x_2, \ldots x_p]$ has the form

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{\frac{-(x-\mu)'\Sigma^{-1}(x-\mu)}{2}} \quad \text{——— ④}$$

where $-\infty \leq x_i \leq \infty \quad i = 1, 2, \ldots p$

we shall denote this p-dimensional normal density by $N_p(\mu, \Sigma)$, which is analogous to the normal density in the univariate case

Properties:

Certain properties of the normal distn will be readed repeatedly in our explanations of statistical models and methods. These properties make it possible to manipulate

normal distn easily.

With the key properties which are responsible for the popularity of the normal distn, the following are true for a random vector x having a multivariate normal distn.

1) Linear combinations of the components of x are normally distributed.

2) All subsets of the components of x have a (multivariate) normal distribution.

3) zero covariance implies that the corresponding components are independently distributed.

4) The conditional distribution of the components are (multivariate) normal.

Bivariate normal density

Let us evaluate the $p = 2$ - variate normal density in terms of the individual parameters $\mu_1 = E(X_1)$, $\mu_2 = E(X_2)$, $\sigma_{11} = \text{Var}(X_1)$, $\sigma_{22} = \text{Var}(X_2)$, and $\rho_{12} = \sigma_{12} / (\sqrt{\sigma_{11}}\sqrt{\sigma_{22}})$

$= \text{Corr}(X_1, X_2)$

We find that the inverse of the covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \qquad \Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

Introducing the correlation co-efficient ρ_{12} by writing

$\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$, we obtain

$\sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \rho_{12}^2)$, and the

squared distance becomes

$$(x-\mu)' \Sigma^{-1} (x-\mu)$$

$$= \begin{bmatrix} x_1-\mu_1, x_2-\mu_2 \end{bmatrix} \frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)} \begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{11} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}$$

$$= \frac{\sigma_{22}(x_1-\mu_1)^2 + \sigma_{11}(x_2-\mu_2)^2 - 2\rho_{12}\sqrt{\sigma_{11}\sigma_{22}}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}$$

$$= \frac{1}{1-\rho_{12}^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right] \quad \text{——— (5)}$$

The last expression is written in terms of the standardised values $(x_1-\mu_1)/\sqrt{\sigma_{11}}$ and $(x_2-\mu_2)/\sqrt{\sigma_{22}}$

Next, since $|\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1-\rho_{12}^2)$ we can substitute for Σ^{-1} and $|\Sigma|$ in (4) to get the expression for the bivariate ($p=2$) normal density involving the individual parameters $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}$ and ρ_{12}

$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}} e^{\frac{-1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]}$$

——— (6)

The expression in (6) is somewhat unwieldy and the compact general form in (4) is more informative in many ways.

On the other hand, the expression in (6) is useful for discussing certain properties of the normal distn. For example,

If the r.v x_1 and x_2 are uncorrelated, so that $\rho_{12} = 0$ then the joint density can be written as the product of 2 univariate normal densities each of the form (1)

That is $f(x_1, x_2) = f(x_1) \cdot f(x_2)$ and x_1 and x_2 are independent

The MLE of mean vector and covariance matrix

The multivariate normal distn is specified completely by the mean vector μ and covariance matrix Σ. The first statistical problem is how to estimate these parameters on the basis of a sample of observations

Given a sample of (vector) observations from a p variate (non degenerate) normal distn we ask for estimators of the mean vector μ and covariance matrix Σ of the distn. The MLE or modifications of them often have some optimum properties

Suppose our sample of n observations on x individual distributed according to $N(\mu, \Sigma)$ is $x_1, x_2, \ldots, x_N$, where $N > p$

The likelihood function is $L = \prod_{\alpha=1}^{N} n(x_\alpha / \mu, \Sigma)$

$$L = \frac{1}{(2\pi)^{\frac{pN}{2}} |\Sigma|^{\frac{1}{2}N}} \exp\left[\frac{1}{2}\sum_{\alpha=1}^{N}(x_\alpha - \mu)'\Sigma^{-1}(x_\alpha - \mu)\right] \quad -(1)$$

In the likelihood function the vectors $x_1, x_2 \cdots x_N$ are fixed at the sample values and L is a function of μ and Σ. To emphasize that these quantities are variables [and not parameters] we shall denote them by μ^* and Σ^*. Then the logarithm of the likelihood function is

$$\log L = -\frac{1}{2}pN\log 2\pi - \frac{1}{2}N\log|\Sigma^*| - \frac{1}{2}\sum_{\alpha=1}^{N}(x_\alpha - \mu^*)' \Sigma^{*-1}(x_\alpha - \mu^*) \quad -(2)$$

Since $\log L$ is an increasing function of L, its maximum is at the same point in the space of μ^*, Σ^* ~~at the same point in the space~~ as the maximum of L. The MLE of μ and Σ are the vector μ^* and the positive definite matrix Σ^* that maximize $\log L$.

Let the sample mean vector be

$$\bar{x} = \frac{1}{N}\sum_{\alpha=1}^{N} x_\alpha = \begin{pmatrix} \frac{1}{N}\sum_{\alpha=1}^{N} x_{1\alpha} \\ \vdots \\ \frac{1}{N}\sum_{\alpha=1}^{N} x_{p\alpha} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} \quad -(3)$$

where $x_\alpha = (x_{1\alpha}, \ldots x_{p\alpha})'$ and $\bar{x}_i = \sum_{\alpha=1}^{N} \frac{x_{i\alpha}}{N}$

Let the matrix of sum of squares and cross products of deviations about the mean be

$$A = \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'$$

$$= \left[\sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j) \right] \qquad i, j = 1, 2, \ldots p$$

and the MLE of ρ_{ij} is $\hat{\rho}_{ij} = \dfrac{\sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)^2} \sqrt{\sum_{\alpha=1}^{N} (x_{j\alpha} - \bar{x}_j)^2}}$