MUTIVARIATE ANALYSIS
(DMSTT24)
(MSC - STATISTICS)

ACHARYA NAGARJUNA UNIVERSITY

CENTRE FOR DISTANCE EDUCATION

NAGARJUNA NAGAR,

GUNTUR

ANDHRA PRADESH

UNIT-I

Concept of Multivariate Analysis:

In botanical and zoological survey the interest of scientists is to investigate the new species of animals or trees and to identify their classes and families. A doctor wants to diagnose the disease of a newly arrived patient. The agricultural scientists investigate the suitability of a piece of land for the cultivation of a particular crop. The selection committee for appointing a group of individuals for a particular job applies different test procedures to select the candidates. Selection committee in educational institute selects students for admission into a course. In all the above cases the investigation is not done on the basis of a single criterion. For example, a doctor only examining the temperature of a patient cannot diagnose it as typhoid. He needs to observe some other symptoms for his proper diagnostic decision. The suitability of apiece of land for a particular crop is judge on test scales of fertility of land, the amount of potash in the land, the soil type, etc. The test scores on educational qualification, age, health condition, behavior, etc., of candidates help the selection committee in selecting a candidate.

The above discussion indicates that a decision regarding an object or individual depends on the simultaneous study of several characteristics observed from the objects. Since the characteristics are observed from an objects, these are inter-related. For a reliable and valid conclusion about any population parameter vector, it needs the study of inter-related variables observed from a sample of objects selected randomly from the population under study. Since the variables cannot be splitted off from each other as these are dependent among themselves.

From the above discussion it is clear that one needs to analyze all the variables observed from n sample objects simultaneously. Multivariate analysis is a statistical technique for simultaneous analysis of two or more variables observed from one or more sample objects. In this analysis, the inter-relationships of the variables are studied along with the

study of mean, variance and some other characteristics related to univariates analysis. However, the main objective of the analysis is to estimate the extent or amount of relationship among the variables. For example, one may need to observe the magnitude and direction of influences of some socio-economic variables which mobilize the couples to adopt family planning method, or to study the causes of preferring a particular industrial item by consumers, or to identify the causes of a disease which affects the community or to identify the class and family of newly observed species. All these analyses depend on multivariate data observed on same occasion from sample individuals.

From the objective of the analysis it is clear that the multivariate data are of two types, viz., dependent and independent. Accordingly, one needs to study the extent of relationship of dependent and independent sets. Also it needs the analysis of the structure of inter-relationships of the variables altogether. Thus multivariate analysis can be classified into two type, viz., (a) Inter-dependence analysis, and (b) Dependent analysis.

The main components of inter-dependence analysis are (i) Principal Component Analysis, (ii) Factor Analysis, (iii) Cluster analysis, where as the dependent analysis deals with (i) discriminant Analysis, (ii)Canonical Correlation Analysis, (iii) Multivariate Analysis of Variance, and (iv) Multivariate Regression Analysis.

Random Vector:

Definition: A random vector is a vector whose elements are random variables i.e., if $X_1, X_2, ..., X_n$ are random variables then the n-tuple

$$\underset{\sim}{\mathbf{X}} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$ is called a random vector.

Distribution function of $\underset{\sim}{\mathbf{X}}$:

The Joint distribution function is $\underset{\sim}{\mathbf{X}}$ is defined by

$$F\left(\underset{\sim}{\mathbf{X}}\right) = F\left(x_1, x_2, ..., x_n\right)$$
$$= P\left(X_1 \le x_1, X_2 \le x_2, ..., X_n \le x_n\right)$$

Which is a monotonic non-decreasing in each of its arguments with basic consideration

$$F\left(-\infty, -\infty, ..., -\infty\right) = 0$$
$$F\left(\infty, \infty, ..., \infty\right) = 1$$

Joint density function of $\underset{\sim}{\mathbf{X}}$:

The Joint probability density function of $\underset{\sim}{\mathbf{x}}$ is defined by

$$f\left(\underset{\sim}{\mathbf{x}}\right) = f\left(X_1, X_2, ..., X_n\right) = \frac{\partial^n F(X_1, X_2, ..., X_n)}{\partial x_1 \partial x_2 ... \partial x_n}$$
$$= \lim_{\Delta x_1 \to 0 ... \Delta x_n \to 0} P\left(x_1 < X_1 < x_1 + \Delta x_1, \cdots, x_n < X_n < x_n + \Delta x_n\right)$$

Conversely, the distribution function can be expressed in terms of the density function as follows

$$F\left(\underset{\sim}{\mathbf{x}}\right) = \int_{-\infty}^{\underset{\sim}{\mathbf{x}}} f\left(\underset{\sim}{\mathbf{x}}\right) \partial \underset{\sim}{\mathbf{x}}$$
$$= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f\left(X_1, X_2, \cdots, X_n\right) \partial X_1 \partial X_2 \cdots \partial X_n$$

Conditional density function:

Definition: The conditional density function of $\underset{\sim}{\mathbf{X}}$ given that the random vector $\underset{\sim}{\mathbf{Y}}$ has a specified value $\underset{\sim}{\mathbf{y}}$ is defined by

$$f\left(\underset{\sim}{\mathbf{x}} / \underset{\sim}{\mathbf{y}}\right) = \lim_{\Delta \underset{\sim}{\mathbf{y}} \to 0} P\left(\underset{\sim}{\mathbf{x}} / \underset{\sim}{\mathbf{y}} \le \underset{\sim}{\mathbf{Y}} \le \underset{\sim}{\mathbf{y}} + \Delta \underset{\sim}{\mathbf{y}}\right)$$
$$= f\left(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n\right) / f\left(y_1, y_2, ..., y_n\right)$$
$$= f\left(\underset{\sim}{\mathbf{x}}, \underset{\sim}{\mathbf{y}}\right) / f\left(\underset{\sim}{\mathbf{y}}\right)$$

where $f\left(\underset{\sim}{\mathbf{y}}\right)$ is the marginal density of $\underset{\sim}{\mathbf{y}}$ and is obtained by the integration.

$$f\left(\underset{\sim}{\mathbf{y}}\right) = \int_{\underset{\sim}{\mathbf{x}}=-\infty}^{\infty} f\left(\underset{\sim}{\mathbf{x}}, \underset{\sim}{\mathbf{y}}\right) \partial \underset{\sim}{\mathbf{x}}$$

Mean vector or Expectation of $\underset{\sim}{\mathbf{X}}$:

The mean vector of is denoted as and is defined as $E(\underset{\sim}{\mathbf{X}})$ and

$$E(\underset{\sim}{\mathbf{X}}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \underset{\sim}{\boldsymbol{\mu}} \text{ (say)}$$

The Variance-Covariance matrix of $\underset{\sim}{\mathbf{X}}$:

The Variance-Covariance matrix of $\underset{\sim}{\mathbf{X}}$ is denoted by $\boldsymbol{\Sigma}$ and is defined as

$$\boldsymbol{\Sigma} = E\left[\left(\underset{\sim}{\mathbf{X}} - E(\underset{\sim}{\mathbf{X}})\right)(\underset{\sim}{\mathbf{X}} - E(\underset{\sim}{\mathbf{X}}))'\right]$$

$$= E[(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}})(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}})']$$

$$= E\left[\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_n - \mu_n \end{bmatrix}\left[X_1 - \mu_1, X_2 - \mu_2, \cdots X_n - \mu_n\right]\right]$$

$$= E\begin{bmatrix} [X_1\text{-}\mu_1]^2 & [(X_1\text{-}\mu_1)(X_2\text{-}\mu_2)] & \dots & [(X_1\text{-}\mu_n)(X_n\text{-}\mu_n)] \\ [(X_2\text{-}\mu_2)(X_1\text{-}\mu_1)] & [X_2\text{-}\mu_2]^2 & \dots & [(X_2\text{-}\mu_2)(X_n\text{-}\mu_n)] \\ \vdots & \vdots & \vdots & \vdots \\ [(X_n\text{-}\mu_n)(X_1\text{-}\mu_1)] & [(X_n\text{-}\mu_n)(X_2\text{-}\mu_2)] & \dots & [X_n\text{-}\mu_n]^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{bmatrix}_{\text{nXn}} \quad \text{where } \sigma_{ij} = E\left[\left(x_i - \mu_i\right)(x_j - \mu_j)\right] = \sigma_{ji}$$

We can see $\boldsymbol{\Sigma}$ is a symmetric matrix

Un-Correlated, Orthogonal and Independent Random Vectors:

Definiton**:** Two random vectors $\underset{\sim}{\mathbf{X}}$ and $\underset{\sim}{\mathbf{Y}}$ are called **Uncorrelated** if

$E\left(\underset{\sim}{\mathbf{X}}\underset{\sim}{\mathbf{Y}}'\right) = E\left(\underset{\sim}{\mathbf{X}}\right)E\left(\underset{\sim}{\mathbf{Y}}'\right)$ (or) $\operatorname{cov}\left(\underset{\sim}{\mathbf{X}}, \underset{\sim}{\mathbf{Y}}\right) = 0$

Orthogonal if $E\left(\underset{\sim}{\mathbf{X}}\underset{\sim}{\mathbf{Y}}'\right)=0$

Independent if $f\left(\underset{\sim}{\mathbf{X}},\underset{\sim}{\mathbf{Y}}\right)=f\left(\underset{\sim}{\mathbf{X}}\right)f\left(\underset{\sim}{\mathbf{Y}}\right)$

Note:

1. If $\underset{\sim}{\mathbf{X}}$ and $\underset{\sim}{\mathbf{Y}}$ are uncorrelated and either $E\left(\underset{\sim}{\mathbf{X}}\right)=0$ (or) $E\left(\underset{\sim}{\mathbf{Y}}\right)=0$ or then $\underset{\sim}{\mathbf{X}}$ and $\underset{\sim}{\mathbf{Y}}$ are orthogonal
2. If $\underset{\sim}{\mathbf{X}}$ and $\underset{\sim}{\mathbf{Y}}$ are independent then

$$E\left(\underset{\sim}{\mathbf{X}}\underset{\sim}{\mathbf{Y}}'\right)=\int_{\underset{\sim}{\mathbf{x}}}\int_{\underset{\sim}{\mathbf{y}}}\underset{\sim}{\mathbf{x}}\underset{\sim}{\mathbf{y}}'f\left(\underset{\sim}{\mathbf{x}}\underset{\sim}{\mathbf{y}}\right)\partial\underset{\sim}{\mathbf{x}}\partial\underset{\sim}{\mathbf{y}}$$

$$=\int_{\underset{\sim}{\mathbf{x}}}\int_{\underset{\sim}{\mathbf{y}}}\underset{\sim}{\mathbf{x}}\underset{\sim}{\mathbf{y}}'f\left(\underset{\sim}{\mathbf{x}}\right)f\left(\underset{\sim}{\mathbf{y}}\right)\partial\underset{\sim}{\mathbf{x}}\partial\underset{\sim}{\mathbf{y}}$$

$$=\left[\int_{\underset{\sim}{\mathbf{x}}}\underset{\sim}{\mathbf{x}}f\left(\underset{\sim}{\mathbf{x}}\right)\partial\underset{\sim}{\mathbf{x}}\right]\left[\int_{\underset{\sim}{\mathbf{y}}}\underset{\sim}{\mathbf{y}}'f\left(\underset{\sim}{\mathbf{y}}\right)\partial\underset{\sim}{\mathbf{y}}\right]=E\left(\underset{\sim}{\mathbf{X}}\right)E\left(\underset{\sim}{\mathbf{Y}}'\right)$$

i.e., and are independent implies uncorrelated

3. The converse of '2' need not be true

MULTIVARIATE NORMAL DISTRIBUTION

MULTIVARIATE NORMAL DENSITY:

Suppose X is a scalar normal variate with mean μ and variance σ^2 then the p.d.f of X can be written as

$$f(x:\mu,\sigma)=ke^{-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)},\sigma^2>0,-\infty<\mu<\infty \qquad \rightarrow (1)$$

Where, k= $\frac{1}{\sigma\sqrt{2\pi}}$

Now suppose $\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$ is a p-variate random vector and

Its mean vector is given by

$$E(\underset{\sim}{\mathbf{X}}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \underset{\sim}{\boldsymbol{\mu}} \qquad \rightarrow (2)$$

and its variance –covariance matrix is given by

$$V(\underset{\sim}{\mathbf{X}}) = E\left[\left(\underset{\sim}{\mathbf{X}} - E(\underset{\sim}{\mathbf{X}})\right)\left(\underset{\sim}{\mathbf{X}} - E(\underset{\sim}{\mathbf{X}})\right)'\right]$$

$$= E[(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}})(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}})']$$

$$V(\underset{\sim}{X}) = \begin{bmatrix} E[X_1-\mu_1]^2 & E[(X_1-\mu_1)(X_2-\mu_2)] & \dots & E[(X_1-\mu_p)(X_p-\mu_p)] \\ E[(X_2-\mu_2)(X_1-\mu_1)] & E[X_2-\mu_2]^2 & \dots & E[(X_2-\mu_2)(X_p-\mu_p)] \\ \vdots & \vdots & \vdots & \vdots \\ E[(X_p-\mu_p)(X_1-\mu_1)] & E[(X_p-\mu_p)(X_2-\mu_2)] & \dots & E[X_p-\mu_p]^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} = \boldsymbol{\Sigma} \text{ (say)} \qquad \rightarrow (3)$$

Where,

$$\sigma_{ij} = E\left[(x_i - \mu_i)(x_j - \mu_j)\right] = \sigma_{ji}$$

clearly, $\boldsymbol{\Sigma}$ is symmetric & positive definite matrix.

Now the multivariate normal density of $\underset{\sim}{\mathbf{X}}$ can be obtained by replacing the positive quantity $(x-\mu)(\sigma^2)^{-1}(x-\mu)$ by the quadratic form

$$(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}) \qquad \rightarrow (4)$$

and is given by

$$f(\underset{\sim}{\mathbf{x}} : \underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = ke^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})} \quad \rightarrow (5)$$

Where (k>0) is chosen so that the integral over the entire p-dimensional

Euclidean space of $X_1, X_2, ..., X_p$ is unity. we observe that

$f(\underset{\sim}{\mathbf{x}} : \underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) \geq 0$ ($\because$ k is chosen as positive)

since $\boldsymbol{\Sigma}$ is positive definite

$$(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}) > 0$$

$$\Rightarrow \quad -\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}) < 0$$

$$\Rightarrow \quad e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})} < e^0 = 1$$

i.e. $0 \leq f(\underset{\sim}{\mathbf{x}} : \underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) \leq k$ i.e. $f(\underset{\sim}{\mathbf{x}})$ is bounded.

Now we should find k(>0) such that

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x) = k \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})' \Sigma^{-1} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})} \mathbf{d}\underset{\sim}{\mathbf{x}} = 1 \quad \rightarrow (6)$$

$$k^{-1} = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})' \Sigma^{-1} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})} \mathbf{d}\underset{\sim}{\mathbf{x}} \quad \rightarrow (7)$$

since $\boldsymbol{\Sigma}^{-1}$ is positive definite $\exists$ a non singular matrix A such that

$$\boldsymbol{\Sigma}^{-1} = A'A \quad \rightarrow (8)$$

then (7) can be written as

$$k^{-1} = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})' A'A (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})} \mathbf{d}\underset{\sim}{\mathbf{x}} \quad \rightarrow (9)$$

If we use the linear transformation from $\underset{\sim}{\mathbf{X}}$ to a new random vector $\underset{\sim}{\mathbf{Y}}$ such that

$$\underset{\sim}{\mathbf{Y}} = A(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}}) \qquad \rightarrow (10)$$

then (9) becomes

$$k^{-1} = J \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}}} \mathbf{d}\underset{\sim}{\mathbf{y}} \qquad \rightarrow (11)$$

where J is the Jacobian obtained when $\underset{\sim}{\mathbf{X}}$ is transformed into $\underset{\sim}{\mathbf{Y}}$ and is given by

$$J^{-1} = \operatorname{mod} \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_p} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial y_2} & \cdots & \frac{\partial y_2}{\partial x_p} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_p}{\partial x_1} & \frac{\partial y_p}{\partial x_2} & \cdots & \frac{\partial y_p}{\partial x_p} \end{vmatrix}$$

$$= \operatorname{mod} \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{vmatrix} = \operatorname{mod}|A|$$

Where $|A|$ is determinant of A.

$\therefore$ Equation (11) becomes

$$k^{-1} = \frac{1}{\operatorname{mod}|A|} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}}} \mathbf{d}\underset{\sim}{\mathbf{y}}$$

$$= \frac{1}{\operatorname{mod}|A|} \prod_{i=1}^{p} \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2} y_i^2} \, dy_i \right)$$

$$= \frac{1}{\operatorname{mod}|A|} \prod_{i=1}^{p} \sqrt{2\pi} \left(\because \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2} y_i^2} \, dy_i = 1 \right)$$

$$= \frac{(2\pi)^{p/2}}{\operatorname{mod}\left|\boldsymbol{\Sigma}^{-1}\right|^{1/2}} \qquad \left(\because \left|\boldsymbol{\Sigma}^{-1}\right| = |A'A| = |A|^2 \right)$$

i.e.,

$$k = \frac{1}{|\boldsymbol{\Sigma}|^{1/2} (2\pi)^{p/2}}$$

($\because \boldsymbol{\Sigma}^{-1}$ is positive definite to $\operatorname{mod}\left|\boldsymbol{\Sigma}^{-1}\right|^{1/2} = \left|\boldsymbol{\Sigma}^{-1}\right|^{1/2} = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}}$)

substituting k in (5) we get the p.d.f of the random normal vector $\underset{\sim}{\mathbf{X}}$ and is given by

$$f(\underset{\sim}{\mathbf{x}} : \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2} (2\pi)^{p/2}} e^{-\frac{1}{2} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})} \qquad \rightarrow (12)$$

thus (12) is the p.d.f of a multivariate normal vector $\underset{\sim}{\mathbf{X}}$ whose mean vector

and variance-covariance matrix are respectively given by $\underset{\sim}{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}$

and is denoted as $n(\underset{\sim}{\mathbf{x}}/\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ and its distribution is denoted as $N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$.

NOTE 1:

From (10) ,we may see that

$$\begin{aligned} E(\underset{\sim}{\mathbf{Y}}) &= AE(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}}) \\ &= A(E(\underset{\sim}{\mathbf{X}}) - \underset{\sim}{\boldsymbol{\mu}}) \\ &= A(\underset{\sim}{\boldsymbol{\mu}} - \underset{\sim}{\boldsymbol{\mu}}) \\ &= \underset{\sim}{\mathbf{0}} \end{aligned}$$

i.e. $\underset{\sim}{\mathbf{Y}}$ has zero mean vector.

The variance-covariance matrix $\underset{\sim}{\mathbf{Y}}$ of is given by

$$V(\underset{\sim}{\mathbf{Y}}) = E[(\underset{\sim}{\mathbf{y}} - E(\underset{\sim}{\mathbf{y}}))(\underset{\sim}{\mathbf{y}} - E(\underset{\sim}{\mathbf{y}}))']$$
$$= E[\underset{\sim}{\mathbf{y}}\underset{\sim}{\mathbf{y}}'] \quad (\because E(\underset{\sim}{\mathbf{y}}) = \underset{\sim}{0})$$
$$= E[A(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\mu})(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\mu})'A']$$
$$= AV(\underset{\sim}{\mathbf{X}})A'$$
$$= A\Sigma A'$$

but from (8),

$$\boldsymbol{\Sigma} = (A'A)^{-1}$$
$$= A^{-1}(A')^{-1} \quad (\because A \text{ is a non-singular})$$
$$\therefore V(\underset{\sim}{\mathbf{Y}}) = AA^{-1}(A')^{-1}A'$$
$$= I_k I_k$$
$$= I_k$$

Thus if $\underset{\sim}{\mathbf{X}}$ is $N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$,then the random vector $\underset{\sim}{\mathbf{Y}}$ defined as $\underset{\sim}{\mathbf{Y}} = A(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}})$ (where A is defined as in (8)) follows $N_p(\underset{\sim}{0}, I_k)$.
In other words , the individual element of $\underset{\sim}{\mathbf{Y}}$ are standard normal variates and mutually independent i.e. $Y_i \sim N(0,1)$ with $\text{cov}(Y_i, Y_j) = 0$.

NOTE 2.
In the practical situations 'A' can be computed as follows .since $\boldsymbol{\Sigma}$ is a symmetric p.d. matrix we may write $\boldsymbol{\Omega}'\boldsymbol{\Sigma}\boldsymbol{\Omega} = \boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_p)$ **,**when $\boldsymbol{\Omega}$ is the normalized latent vector matrix and $\boldsymbol{\Lambda}$ is the latent root matrix and since $\boldsymbol{\Sigma}$ is p.d. all $\lambda_1, \lambda_2, ..., \lambda_p$ are positive .Therefore $\boldsymbol{\Lambda}$ can be written $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}^{1/2})'(\boldsymbol{\Lambda}^{1/2})$
Where, $\boldsymbol{\Lambda}^{1/2} = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_p})$
then

$$\boldsymbol{\Omega}'\boldsymbol{\Sigma}\boldsymbol{\Omega} = (\boldsymbol{\Lambda}^{1/2})'(\boldsymbol{\Lambda}^{1/2})$$
$$\Rightarrow \boldsymbol{\Sigma} = (\boldsymbol{\Omega}^{-1})'(\boldsymbol{\Lambda}^{1/2})'\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Omega}^{-1} = A^{-1}(A^{-1})'$$

where

$$A^{-1} = (\boldsymbol{\Omega}^{-1})'(\boldsymbol{\Lambda}^{1/2})'$$
$$\Rightarrow A = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Omega}' \quad (\because (\Lambda^{1/2})' = \Lambda^{1/2})$$
thus, $\underset{\sim}{\mathbf{Y}} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Omega}'(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}})$
where $\boldsymbol{\Lambda}^{1/2} = diag(\frac{1}{\sqrt{\lambda_1}}, ..., \frac{1}{\sqrt{\lambda_p}})$

The transformation from $\underset{\sim}{\mathbf{X}}$ to $\underset{\sim}{\mathbf{Y}}$ follows $N_p(\underset{\sim}{\mathbf{0}}, \mathbf{I}_k)$ .This transformation is called "whitening".

ALTERNATIVE METHOD OF OBTAINING THE P.D.F. OF MULTIVARIATE NORMAL VARIATE:-

Suppose $\underset{\sim}{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix}$

is a multivariate normal vector with mean vector $\underset{\sim}{0}$ and variance-covariance matrix I_p .

i.e. $E(\underset{\sim}{\mathbf{Y}}) = \underset{\sim}{\mathbf{0}}$

i.e. $E(y_i) = 0 \forall i$ → (a)

and

$$\mathrm{V}(\underset{\sim}{\mathbf{Y}}) = \begin{bmatrix} E(y_1^2) & E(y_1 y_2) & \dots & E(y_1 y_p) \\ E(y_2 y_1) & E(y_2^2) & \dots & E(y_2 y_p) \\ \vdots & \vdots & \dots & \vdots \\ E(y_p y_1) & E(y_p y_2) & \dots & E(y_p^2) \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = I_p$$

i.e. $E(Y_i^2) = 1$ & $E(Y_i Y_j) = 0 \quad \forall i \neq j$ → **(b)**

Thus $Y_i's$ are i.i.d. with '0' mean and unit variance .now consider the p.d.f. of $\underset{\sim}{\mathbf{Y}}$,which is given as

$$f(\underset{\sim}{\mathbf{Y}}) = f(y_1, y_2, \dots, y_p)$$
$$= f(y_1) f(y_2) \dots f(y_p)$$
$$= \prod_{i=1}^{p} f(y_i) \quad \text{(since } y_i's \text{ are i.i.d.)}$$

$$= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_i^2} \qquad (\text{since } Y_i \sim N(0,1)),\ f(y_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_i^2})$$

$$= \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\sum_{j=1}^{n} y_i^2} \qquad = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}}} \qquad \rightarrow (1)$$

let us make use of the following linear transformation of $\underset{\sim}{\mathbf{Y}}$ into $\underset{\sim}{\mathbf{X}}$

$\underset{\sim}{\mathbf{X}} = \mathrm{A}\underset{\sim}{\mathbf{Y}} + \underset{\sim}{\mathbf{b}}$ where A is non-singular $\rightarrow (2)$

Now the p.d.f of $\underset{\sim}{\mathbf{X}}$ becomes

$$f(\underset{\sim}{\mathbf{x}}) = f(x_1, x_2, ..., x_p)$$

$$= \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\mathbf{b}})'(A^{-1})'A^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\mathbf{b}})} J(\underset{\sim}{\mathbf{x}})$$

where $J(\underset{\sim}{\mathbf{x}})$ is the jacobian and is given by

$$J(\underset{\sim}{\mathbf{x}}) = \left| \frac{d\underset{\sim}{\mathbf{y}}}{d\underset{\sim}{\mathbf{x}}} \right|$$

$$= \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_p} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_p} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_p}{\partial x_1} & \frac{\partial y_p}{\partial x_2} & \cdots & \frac{\partial y_p}{\partial x_p} \end{vmatrix}$$

$$= \left| A^{-1} \right|$$

$$= \frac{1}{|A|} \qquad (\because \left| A^{-1} \right| = \frac{1}{|A|})$$

$$\therefore f(\underset{\sim}{\mathbf{x}}) = \frac{1}{|A|(2\pi)^{p/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\mathbf{b}})'(AA')^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\mathbf{b}})}$$

Which is the p.d.f. of $\underset{\sim}{\mathbf{X}}$ in terms of A & $\underset{\sim}{\mathbf{b}}$ and now let us interpret $\underset{\sim}{\mathbf{b}}$ and A
From (2),

$$E(\underset{\sim}{\mathbf{X}}) = AE(\underset{\sim}{\mathbf{Y}}) + \underset{\sim}{\mathbf{b}} = \underset{\sim}{\mathbf{b}} \qquad (\because E(\underset{\sim}{\mathbf{Y}}) = \underset{\sim}{\mathbf{0}})$$

i.e $\underset{\sim}{\mathbf{b}} = E(\underset{\sim}{\mathbf{X}}) = \underset{\sim}{\boldsymbol{\mu}}$ say → (4)

$$\mathrm{V}(\underset{\sim}{\mathbf{X}}) = \mathrm{AV}(\underset{\sim}{\mathbf{Y}})'A' = AI_pA' = AA' \quad \text{(from (b))}$$
$$= \boldsymbol{\Sigma} \text{ say} \qquad \rightarrow (5)$$

$\therefore \underset{\sim}{\mathbf{b}}$ is the mean vector of $\underset{\sim}{\mathbf{X}}$ and AA' is the variance-covariance matrix of $\underset{\sim}{\mathbf{X}}$

using (4) & (5) in (3) we get,

$$f(\underset{\sim}{\mathbf{x}}) = \frac{1}{(2\pi)^{p/2} \bmod |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})} \qquad \rightarrow (6)$$

$$(\because |AA'| = |\boldsymbol{\Sigma}| \quad \Rightarrow |A|^2 = |\boldsymbol{\Sigma}| \quad \Rightarrow |A| = |\boldsymbol{\Sigma}|^{1/2})$$

since $|\boldsymbol{\Sigma}| = |A|^2 > 0, \boldsymbol{\Sigma}$ is positive definite and therefore $\bmod |\boldsymbol{\Sigma}|^{1/2} = |\boldsymbol{\Sigma}|^{1/2}$ and (6) can be written as

$$f(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})} \qquad \rightarrow (7)$$

Eq (7) is the p.d.f . of the multivariate normal variate $\underset{\sim}{\mathbf{X}}$ where mean is $\underset{\sim}{\boldsymbol{\mu}}$ and variance –covariance matrix is $\boldsymbol{\Sigma}$ and is denoted by $n(\underset{\sim}{\mathbf{x}}/\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$.
The distribution function of $\underset{\sim}{\mathbf{X}}$ is denoted as $N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$.

Definition: If $\underset{\sim}{\mathbf{X}}$ is a random vector with mean $\underset{\sim}{\boldsymbol{\mu}}$ and the variance-covariance matrix , $\boldsymbol{\Sigma}$ and its p.d.f. is given by (7) then $\underset{\sim}{\mathbf{X}}$ is said to follows p-variate normal distribution and is denoted as $\underset{\sim}{\mathbf{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \Sigma)$.

PROFPERTIES OF MULTIVARIATE NORMAL DISTRIBUTION

THEOREM:

Let $\underset{\sim}{\mathbf{X}}$ (with p components) be distributed according to $N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ then, $\underset{\sim}{\mathbf{Y}} = C\underset{\sim}{\mathbf{X}}$ is distributed according to $N(C\underset{\sim}{\boldsymbol{\mu}}, C\boldsymbol{\Sigma}C')$ for C non-singular.

PROOF:

Since $\underset{\sim}{\mathbf{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ & its p.d.f. is given as

$$f(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})} \quad \rightarrow (1)$$

Now ,consider the linear transformation

$\underset{\sim}{\mathbf{Y}} = C\underset{\sim}{\mathbf{X}}$ where C is non-singular

$$\Rightarrow \underset{\sim}{\mathbf{X}} = C^{-1}\underset{\sim}{\mathbf{Y}} \quad \rightarrow (2)$$

Now the p.d.f. 1 becomes interms of $\underset{\sim}{\mathbf{Y}}$ as

$$g(\underset{\sim}{\mathbf{y}}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(C^{-1}\underset{\sim}{\mathbf{y}} - \underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(C^{-1}\underset{\sim}{\mathbf{y}} - \underset{\sim}{\boldsymbol{\mu}})} J(\underset{\sim}{\mathbf{y}}) \quad \rightarrow (3)$$

where $J(\underset{\sim}{\mathbf{y}})$ is the Jacobian and is given by

$$\begin{aligned} J(\underset{\sim}{\mathbf{y}}) &= \text{mod}\left|\frac{\partial \underset{\sim}{\mathbf{x}}}{\partial \underset{\sim}{\mathbf{y}}}\right| = \text{mod}\left|C^{-1}\right| \\ &= \frac{1}{\text{mod}|C|} \\ &= \frac{1}{\sqrt{|C|^2}} \\ &= \sqrt{\frac{|\boldsymbol{\Sigma}|}{|C||\boldsymbol{\Sigma}||C'|}} \\ &= \frac{|\boldsymbol{\Sigma}|^{1/2}}{|C\boldsymbol{\Sigma}C'|^{1/2}} \quad \rightarrow (4) \end{aligned}$$

Using (4) & (3) ,we get

$$g(\underset{\sim}{\mathbf{y}}) = \frac{1}{(2\pi)^{p/2}\left|C\boldsymbol{\Sigma}C'\right|^{1/2}} e^{-\frac{1}{2}(C^{-1}\underset{\sim}{\mathbf{y}} - \underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(C^{-1}\underset{\sim}{\mathbf{y}} - \underset{\sim}{\boldsymbol{\mu}})}$$

$$= \frac{1}{(2\pi)^{p/2}\left|C\boldsymbol{\Sigma}C'\right|^{1/2}} e^{-\frac{1}{2}[C^{-1}(\underset{\sim}{\mathbf{y}} - C\underset{\sim}{\boldsymbol{\mu}})]'\boldsymbol{\Sigma}^{-1}[C^{-1}(\underset{\sim}{\mathbf{y}} - C\underset{\sim}{\boldsymbol{\mu}})]}$$

$$= \frac{1}{(2\pi)^{p/2}\left|C\boldsymbol{\Sigma}C'\right|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{y}} - C\underset{\sim}{\boldsymbol{\mu}})'(C\boldsymbol{\Sigma}C')^{-1}(\underset{\sim}{\mathbf{y}} - C\underset{\sim}{\boldsymbol{\mu}})}$$

$$= n(\underset{\sim}{\mathbf{y}} / C\underset{\sim}{\boldsymbol{\mu}}, C\boldsymbol{\Sigma}C') \quad \rightarrow (5)$$

But

$$E(\underset{\sim}{\mathbf{Y}}) = CE(\underset{\sim}{\mathbf{X}}) = C\underset{\sim}{\boldsymbol{\mu}} \quad \rightarrow (6)$$

$$\& \ \mathrm{V}(\underset{\sim}{\mathbf{Y}}) = C\mathrm{V}(\underset{\sim}{\mathbf{X}})C' = C\sum C' \quad \rightarrow (7)$$

Now, if we write the multivariate normal p.d.f. of $\underset{\sim}{\mathbf{Y}}$ with mean $\underset{\sim}{\boldsymbol{\mu}}$ and the variance-covariance matrix $C\boldsymbol{\Sigma}C'$ that will becomes as (5) and therefore

$$C\underset{\sim}{\mathbf{X}} \sim N(C\underset{\sim}{\boldsymbol{\mu}}, C\boldsymbol{\Sigma}C').$$

Hence the proof.

THEOREM:

If a multivariate normal vector is divided into two subvectors and one sub -vector is uncorrelated with other sub-vector ,then those two sub-vectors of variables are independent and each sub-vector is also a multivariate normal vector.

(OR)

Let $\underset{\sim}{\mathbf{X}}_{p\times 1} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \Sigma) \ \& \ \underset{\sim}{\mathbf{X}} = \begin{pmatrix} \underset{\sim}{\mathbf{X}}_1 \\ \underset{\sim}{\mathbf{X}}_2 \end{pmatrix}$

Where $\underset{\sim}{\mathbf{X}}_1$ **is** $q\times 1$ **and** $\underset{\sim}{\mathbf{X}}_2$ **is** $(p-q)\times 1$ **and** $\underset{\sim}{\boldsymbol{\mu}} = \begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}}_1 \\ \underset{\sim}{\boldsymbol{\mu}}_2 \end{pmatrix}$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where , Σ_{11} **is variance-covariance matrix of** $\underset{\sim}{X}_1$

Σ_{22} **is variance-covariance matrix of** $\underset{\sim}{X}_2$

and Σ_{12} **is covariance matrix of** $\underset{\sim}{X}_1 \& \underset{\sim}{X}_2$

Σ_{21} **is covariance matrix of** $\underset{\sim}{X}_2 \& \underset{\sim}{X}_1$.

Now if $\Sigma_{12} = \Sigma_{21}' = 0_{q \times p-q}$

then , $\underset{\sim}{X}_2 \& \underset{\sim}{X}_1$ **are independent and** $\underset{\sim}{X}_1 \sim N_q(\underset{\sim}{\mu}_1, \Sigma_{11})$, $\underset{\sim}{X}_2 \sim N_{p-q}(\underset{\sim}{\mu}_2, \Sigma_{22})$

PROOF:

We are given $\Sigma_{12} = \mathbf{0}_{qXpq} = \Sigma_{21}'$

i.e. the covariance matrix of $\underset{\sim}{X}_{p \times 1}$ is given by

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

In order to show that, the random vectors $\underset{\sim}{X}_1$ **&** $\underset{\sim}{X}_2$ are independently normally distributed, we have to show that

$$n(\underset{\sim}{x}/\underset{\sim}{\mu}, \Sigma) = n(\underset{\sim}{x}_1/\underset{\sim}{\mu}_1, \Sigma_{11}) n(\underset{\sim}{x}_2/\underset{\sim}{\mu}_2, \Sigma_{22})$$

we have,

$$n(\underset{\sim}{x}/\underset{\sim}{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu})} \quad \rightarrow (1)$$

consider the Q.F in (1),

i.e., $Q = (\underset{\sim}{x} - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu})$

$$= \begin{bmatrix} \underset{\sim}{x}_1 - \underset{\sim}{\mu}_1 \\ \underset{\sim}{x}_2 - \underset{\sim}{\mu}_2 \end{bmatrix}' \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \underset{\sim}{x}_1 - \underset{\sim}{\mu}_1 \\ \underset{\sim}{x}_2 - \underset{\sim}{\mu}_2 \end{bmatrix}$$

$$= \begin{bmatrix} (\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1)' & (\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)' \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} \underset{\sim}{x}_1 - \underset{\sim}{\mu}_1 \\ \underset{\sim}{x}_2 - \underset{\sim}{\mu}_2 \end{bmatrix}$$

($\because \Sigma_{11}$ is the variance-covariance matrix of $\underset{\sim}{X}_1$ and hence positive definite)

$$= \begin{bmatrix} (\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1)'\Sigma_{11}^{-1} & (\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)'\Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} \underset{\sim}{x}_1 - \underset{\sim}{\mu}_1 \\ \underset{\sim}{x}_2 - \underset{\sim}{\mu}_2 \end{bmatrix}$$

$$= (\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1)'\Sigma_{11}^{-1}(\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1) + (\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)'\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)$$

$$= Q_1 + Q_2 \qquad \rightarrow (2)$$

also we have,

$$|\Sigma| = \begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix} = |\Sigma_{11}||\Sigma_{22}| \qquad \rightarrow (3)$$

Using (2) & (3) in (1) we get ,

$$n(\underset{\sim}{x}/\underset{\sim}{\mu}, \Sigma) = \frac{1}{(2\pi)^{q/2}|\Sigma_{11}|^{1/2}} e^{-\frac{1}{2}Q_1} \frac{1}{(2\pi)^{(p-q)/2}|\Sigma_{22}|^{1/2}} e^{-\frac{1}{2}Q_2}$$

where $Q_1 \& Q_2$ are is as given in (2),

$$\therefore n(\underset{\sim}{x}/\underset{\sim}{\mu}, \Sigma) = n(\underset{\sim}{x}_1/\underset{\sim}{\mu}_1, \Sigma_{11}).n(\underset{\sim}{x}_2/\underset{\sim}{\mu}_2, \Sigma_{22})$$

Thus ,the joint p.d.f. of the normal variates $X_1, X_2, ..., X_p$ is the product of the marginal p.d.f. of $X_1, X_2, ..., X_q$ and the marginal p.d.f. of $X_{q+1}, ..., X_p$. Thus, the two sets of normal variates are independent.

THEOREM:

If $\underset{\sim}{X}_1 \& \underset{\sim}{X}_2$ are independent and are distributed as $N_q(\underset{\sim}{\mu}_1, \Sigma_{11}) \& N_{p-q}(\underset{\sim}{\mu}_2, \Sigma_{22})$ respectively then , $\begin{pmatrix} \underset{\sim}{X}_1 \\ \underset{\sim}{X}_2 \end{pmatrix} \sim N_p\left(\begin{bmatrix} \underset{\sim}{\mu}_1 \\ \underset{\sim}{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \right)$.

PROOF:-

we have given ,

$$\underset{\sim}{X}_1 \sim N_q(\underset{\sim}{\mu}_1, \Sigma_{11})$$

$$\underset{\sim}{X}_2 \sim N_{p-q}(\underset{\sim}{\mu}_2, \Sigma_{22})$$

and $\underset{\sim}{X}_1 \& \underset{\sim}{X}_2$ are independent i.e. $\underset{\sim}{X}_1, \underset{\sim}{X}_2$ are uncorrelated.

i.e. $\text{cov}(\underset{\sim}{X}_1, \underset{\sim}{X}_2) = \Sigma_{12} = \Sigma_{21} = 0$.

We have to find out the joint p.d.f. of $f(\underset{\sim}{\mathbf{x}})$ of $\underset{\sim}{\mathbf{X}} = \begin{pmatrix} \underset{\sim}{\mathbf{X}}_1 \\ \underset{\sim}{\mathbf{X}}_2 \end{pmatrix}$

we have,

$g(\underset{\sim}{\mathbf{x}}) = f(\underset{\sim}{\mathbf{x}}_1) f(\underset{\sim}{\mathbf{x}}_2)$ ($\because \underset{\sim}{\mathbf{X}}_1 \& \underset{\sim}{\mathbf{X}}_2$ are independent)

$$= n(\underset{\sim}{\mathbf{x}}_1 / \underset{\sim}{\boldsymbol{\mu}}_1, \boldsymbol{\Sigma}_{11}).n(\underset{\sim}{\mathbf{x}}_2 / \underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}_{22})$$

$$= \frac{1}{(2\pi)^{q/2} |\boldsymbol{\Sigma}_{11}|^{1/2}} e^{-\frac{1}{2}Q_1} \frac{1}{(2\pi)^{(p-q)/2} |\boldsymbol{\Sigma}_{22}|^{1/2}} e^{-\frac{1}{2}Q_2}, \text{ (where } Q_i = (\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}_i)' \boldsymbol{\Sigma}_{ii}^{-1} (\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}_i), \text{ i=1,2)}$$

$$= \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(Q_1+Q_2)} \quad \left(\because |\boldsymbol{\Sigma}| = \begin{vmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{vmatrix} = |\boldsymbol{\Sigma}_{11}||\boldsymbol{\Sigma}_{22}| \right) \quad \rightarrow (1)$$

$$\text{Where } Q_1 + Q_2 = (\underset{\sim}{\mathbf{x}}_1 - \underset{\sim}{\boldsymbol{\mu}}_1)' \boldsymbol{\Sigma}_{11}^{-1} (\underset{\sim}{\mathbf{x}}_1 - \underset{\sim}{\boldsymbol{\mu}}_1) + (\underset{\sim}{\mathbf{x}}_2 - \underset{\sim}{\boldsymbol{\mu}}_2)' \boldsymbol{\Sigma}_{22}^{-1} (\underset{\sim}{\mathbf{x}}_2 - \underset{\sim}{\boldsymbol{\mu}}_2) \quad \rightarrow (2)$$

Let us consider

$$Q = (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}) \quad \rightarrow (3)$$

where , $\underset{\sim}{\mathbf{x}} = \begin{pmatrix} \underset{\sim}{\mathbf{x}}_1 \\ \underset{\sim}{\mathbf{x}}_2 \end{pmatrix}, \underset{\sim}{\boldsymbol{\mu}} = \begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}}_1 \\ \underset{\sim}{\boldsymbol{\mu}}_2 \end{pmatrix}$ is $E(\underset{\sim}{\mathbf{X}})$ and the variance-covariance matrix $\underset{\sim}{\mathbf{X}}$ is

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{V}(\underset{\sim}{\mathbf{X}}_1) & \mathrm{cov}(\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2) \\ \mathrm{cov}(\underset{\sim}{\mathbf{X}}_2, \underset{\sim}{\mathbf{X}}_1) & \mathrm{V}(\underset{\sim}{\mathbf{X}}_2) \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

But ,since $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21} = \mathbf{0}_{qXp-q}$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{pmatrix} \quad \rightarrow (4)$$

Now, $Q = \begin{pmatrix} \underset{\sim}{\mathbf{x}}_1 - \underset{\sim}{\boldsymbol{\mu}}_1 \\ \underset{\sim}{\mathbf{x}}_2 - \underset{\sim}{\boldsymbol{\mu}}_2 \end{pmatrix}' \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix} \begin{pmatrix} \underset{\sim}{\mathbf{x}}_1 - \underset{\sim}{\boldsymbol{\mu}}_1 \\ \underset{\sim}{\mathbf{x}}_2 - \underset{\sim}{\boldsymbol{\mu}}_2 \end{pmatrix}$

$$= (\underset{\sim}{\mathbf{x}}_1 - \underset{\sim}{\boldsymbol{\mu}}_1)' \boldsymbol{\Sigma}_{11}^{-1} (\underset{\sim}{\mathbf{x}}_1 - \underset{\sim}{\boldsymbol{\mu}}_1) + (\underset{\sim}{\mathbf{x}}_2 - \underset{\sim}{\boldsymbol{\mu}}_2)' \boldsymbol{\Sigma}_{22}^{-1} (\underset{\sim}{\mathbf{x}}_2 - \underset{\sim}{\boldsymbol{\mu}}_2) \quad \rightarrow (5)$$

from (2) & (5) , $Q_1 + Q_2 = Q$

$$\therefore g(\underset{\sim}{\mathbf{X}}) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-\frac{1}{2}Q}$$

Where, Q is given by (3) but $g(\underset{\sim}{\mathbf{X}})$ is nothing but $n(\underset{\sim}{\mathbf{x}}/\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$.

Thus $\underset{\sim}{\mathbf{X}} = \begin{pmatrix} \underset{\sim}{\mathbf{X_1}} \\ \underset{\sim}{\mathbf{X_2}} \end{pmatrix} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ where, $\boldsymbol{\Sigma}$ is as given by (4).

THEOREM:

If $X_1, X_2, ..., X_p$ have a joint normal distribution , a necessary & sufficient condition for one subset of some random variables and the subset consisting of the remaining random variables be independent is that each covariance of a variable from one set and a variable from the other set be '0'.

PROOF:-

Necessary condition:-

With out loss of generality let us assume that the first q variables form the first subset and the remaining p-q variables form the second subset.

In order to prove the necessary condition, we have given that the variables of $X_1, X_2, ..., X_q$ are independently distributed with the variables $X_{q+1}, X_{q+2}, ..., X_p$ and we have to prove

$$\text{cov}(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))] = 0$$

where , $1 \le i \le q \;\&\; q+1 \le j \le p$

we have

$$\text{cov}(X_i, X_j) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} (x_i - E(X_i)(x_j - E(X_j)) f(x_1, x_2, ..., x_p) dx_1 dx_2 ... dx_p$$

$$= \left(\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} (x_i - E(X_i)) f_1(x_1 ... x_q) dx_1 ... dx_q \right).$$

$$\left(\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} (x_j - E(X_j)) f_2(x_{q+1} ... x_p) dx_{q+1} ... dx_p \right)$$

$$\left(\because f(x_1, ..., x_p) = f_1(x_1, ..., x_q) f_2(x_{q+1}, ..., x_p) \right)$$

$$= E[X_i - E(X_i)] E[X_j - E(X_j)]$$

$$= [E(X_i) - E(X_i)][E(X_j) - E(X_j)]$$

$$= 0.0$$

$$= 0$$

Thus if one set of variables is independent of the remaining variables then, the set of variables are uncorrelated with the other set of variables.

SUFFICIENT CONDITION:

Here we have given

$$\underset{\sim}{\mathbf{X}} = \begin{pmatrix} \underset{\sim}{\mathbf{X_1}} \\ \underset{\sim}{\mathbf{X_2}} \end{pmatrix} \& \ \underset{\sim}{\mathbf{X}} : \mathbf{N}(\underset{\sim}{\boldsymbol{\mu}}, \Sigma) \qquad \& \operatorname{cov}(X_i, X_j) = 0$$

where , X_i is from $\underset{\sim}{\mathbf{X_1}}$

X_j is from $\underset{\sim}{\mathbf{X_2}}$

i.e. $\operatorname{cov}(\underset{\sim}{\mathbf{X_1}}, \underset{\sim}{\mathbf{X_2}}) = \Sigma_{12} = 0_{q \times p-q}$ and we have to prove $\underset{\sim}{\mathbf{X_1}}$ **&** $\underset{\sim}{\mathbf{X}}_2$ are independently distributed.

The proof is already available.

NOTE:

To prove the necessary condition of the above theorem we need not assume $X_1, ..., X_p$ are normally distributed.

THEOREM:

If $\underset{\sim}{\mathbf{X}} \sim \mathbf{N_p}(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ and if a set of components of $\underset{\sim}{\mathbf{X}}$ is uncorrelated with the set of other components, the marginal distribution of the set is multivariate normal with means, variances and co-variances obtained by taking the proper components of $\underset{\sim}{\boldsymbol{\mu}}$ and Σ respectively.

PROOF:

Without loss of generality let us assume that the set consists of first 'q' components of $\underset{\sim}{\mathbf{X}}$ is uncorrelated with other components.

i.e. if $\underset{\sim}{\mathbf{X}} = \begin{pmatrix} \underset{\sim}{\mathbf{X_1}} \\ \underset{\sim}{\mathbf{X_2}} \end{pmatrix}$ where $\underset{\sim}{\mathbf{X}}_1$ is $q \times 1$ & $\underset{\sim}{\mathbf{X_2}}$ is $(p-q) \times 1$

$\operatorname{cov}(\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2) = 0 = \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21}$

i.e. $\underset{\sim}{\mathbf{X}}_1$ **&** $\underset{\sim}{\mathbf{X}}_2$ are independent.

What all we have to prove is the marginal distribution of $\underset{\sim}{\mathbf{X}}_1$ is $N_q(\underset{\sim}{\boldsymbol{\mu}}_1, \boldsymbol{\Sigma}_{11})$.

This is already proved above.

THEOREM:

If $\underset{\sim}{\mathbf{X}}$ has multivariate normal distribution, then any subset of the components of $\underset{\sim}{\mathbf{X}}$ have a (multivariate) normal distribution.

(OR)

If $\underset{\sim}{\mathbf{X}}$ is distributed as $N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, the marginal distribution of any (sub) set of components of $\underset{\sim}{\mathbf{X}}$ is multivariate normal with means, variances and co-variances obtained by taking the proper components of $\underset{\sim}{\boldsymbol{\mu}}$ and Σ respectively.

PROOF:

Let $\underset{\sim}{\mathbf{X}} = \begin{pmatrix} \underset{\sim}{\mathbf{X_1}} \\ \underset{\sim}{\mathbf{X_2}} \end{pmatrix}, \underset{\sim}{\boldsymbol{\mu}} = \begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}_1} \\ \underset{\sim}{\boldsymbol{\mu}_2} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$

where

$$\underset{\sim}{\boldsymbol{\mu}}_1 = E(\underset{\sim}{\mathbf{X}}_1), \underset{\sim}{\boldsymbol{\mu}}_2 = E(\underset{\sim}{\mathbf{X}}_2)$$

$$\boldsymbol{\Sigma}_{11} = \mathrm{V}(\underset{\sim}{\mathbf{X}}_1), \boldsymbol{\Sigma}_{22} = \mathrm{V}(\underset{\sim}{\mathbf{X}}_2)$$

$$\boldsymbol{\Sigma}_{12} = \mathrm{cov}(\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2') = \boldsymbol{\Sigma}_{21}' = \left[\mathrm{cov}(\underset{\sim}{\mathbf{X}}_2, \underset{\sim}{\mathbf{X}}_1)\right]'$$

Now we shall make a non singular linear transformation to sub vectors

$$\underset{\sim}{\mathbf{Y}}_1 = \underset{\sim}{\mathbf{X}}_1 + \mathbf{M}\underset{\sim}{\mathbf{X}}_2$$
$$\underset{\sim}{\mathbf{Y}}_2 = \underset{\sim}{\mathbf{X}}_2 \qquad \rightarrow (1)$$

choosing M so that the components of $\underset{\sim}{\mathbf{Y}}_1$ are uncorrelated with the components of $\underset{\sim}{\mathbf{Y}}_2 = \underset{\sim}{\mathbf{X}}_2$. The matrix 'M' must satisfy the equation.

$$\begin{aligned}
\mathrm{cov}(\underset{\sim}{\mathbf{Y}}_1, \underset{\sim}{\mathbf{Y}}_2) = \mathbf{0}_{qXp-q} &= \mathrm{E}\left[(\underset{\sim}{\mathbf{Y}}_1 - \mathrm{E}(\underset{\sim}{\mathbf{Y}}_1))(\underset{\sim}{\mathbf{Y}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{Y}}_2))'\right] \\
&= \mathrm{E}\left[\left\{(\underset{\sim}{\mathbf{X}}_1 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_1)) + \mathrm{M}(\underset{\sim}{\mathbf{X}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_2))'\right\}\left\{\underset{\sim}{\mathbf{X}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_2)\right\}'\right] \\
&= \mathrm{E}\left\{\left\{\underset{\sim}{\mathbf{X}}_1 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_1)\right\}\left\{\underset{\sim}{\mathbf{X}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_2)\right\}'\right\} + \mathrm{ME}\left\{\left\{(\underset{\sim}{\mathbf{X}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_2))\right\}\left\{\underset{\sim}{\mathbf{X}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_2)\right\}'\right\} \\
&= \mathrm{cov}(\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2') + \mathrm{M}\,\mathrm{cov}(\underset{\sim}{\mathbf{X}}_2, \underset{\sim}{\mathbf{X}}_2') \\
&= \boldsymbol{\Sigma}_{12} + \mathrm{M}\boldsymbol{\Sigma}_{22}
\end{aligned}$$

Thus, $\mathrm{M} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ → (2)

and the vector $\underset{\sim}{\mathbf{Y}}_1$ becomes

$$\underset{\sim}{\mathbf{Y}}_1 = \underset{\sim}{\mathbf{X}}_1 - \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{22}^{-1}\underset{\sim}{\mathbf{X}}_2 \qquad \rightarrow (3)$$

and the vector

$$\underset{\sim}{\mathbf{Y}} = \begin{pmatrix} \underset{\sim}{\mathbf{Y}}_1 \\ \underset{\sim}{\mathbf{Y}}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\mathbf{X}}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\underset{\sim}{\mathbf{X}}_2 \\ \underset{\sim}{\mathbf{X}}_2 \end{pmatrix}$$

$$= \begin{pmatrix} I_q & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0}_{p-qXq} & I_{p-q} \end{pmatrix} \begin{pmatrix} \underset{\sim}{\mathbf{X}}_1 \\ \underset{\sim}{\mathbf{X}}_2 \end{pmatrix} = \begin{pmatrix} I_q & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ 0 & I_{p-q} \end{pmatrix} \underset{\sim}{\mathbf{X}}$$

$$= C\underset{\sim}{X}$$

Since C is a non singular matrix, $\underset{\sim}{\mathbf{Y}}$ is non-singular transformation of $\underset{\sim}{\mathbf{X}}$ and therefore has a normal distribution with mean vector given by

$$\underset{\sim}{\mathbf{v}} = \begin{pmatrix} I_q & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ 0 & I_{p-q} \end{pmatrix} \begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}}_1 \\ \underset{\sim}{\boldsymbol{\mu}}_2 \end{pmatrix}$$

$$= \begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\underset{\sim}{\boldsymbol{\mu}}_2 \\ \underset{\sim}{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\mathbf{v}}_1 \\ \underset{\sim}{\mathbf{v}}_2 \end{pmatrix} \qquad \text{(say)}$$

and the variance –covariance matrix is given by

$$\Omega = \mathrm{V}(\underset{\sim}{\mathbf{Y}}) = \begin{pmatrix} \mathrm{V}(\underset{\sim}{\mathbf{y}}_1) & \mathrm{cov}(\underset{\sim}{\mathbf{y}}_1, \underset{\sim}{\mathbf{y}}_2) \\ \mathrm{cov}(\underset{\sim}{\mathbf{y}}_2, \underset{\sim}{\mathbf{y}}_1) & \mathrm{V}(\underset{\sim}{\mathbf{y}}_2) \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22} \\ \boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

which implies that $\underset{\sim}{\mathbf{Y}}_1 \,\&\, \underset{\sim}{\mathbf{Y}}_2$ are uncorrelated and further $\underset{\sim}{\mathbf{Y}} = \begin{pmatrix} \underset{\sim}{\mathbf{Y}}_1 \\ \underset{\sim}{\mathbf{Y}}_2 \end{pmatrix}$ has multivariate normal distribution.

Therefore $\underset{\sim}{\mathbf{Y}}_1 \,\&\, \underset{\sim}{\mathbf{Y}}_2$ are independent.

$\therefore \underset{\sim}{\mathbf{Y}}_2 = \underset{\sim}{\mathbf{X}}_2$ has the marginal distribution $N_{p-q}(\underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}_{22})$.

CONDITIONAL DISTRIBUTION:-

In the above we have seen $\underset{\sim}{\mathbf{Y}}_1 = \underset{\sim}{\mathbf{X}}_1 - \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mathbf{X}}_2 \,\&\, \underset{\sim}{\mathbf{Y}}_2 = \underset{\sim}{\mathbf{X}}_2$ are uncorrelated and therefore they are independently distributed.

Write down,

$$\underset{\sim}{\mathbf{Y}} = \begin{pmatrix} \underset{\sim}{\mathbf{Y}}_1 \\ \underset{\sim}{\mathbf{Y}}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\mathbf{X}}_1 - \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mathbf{X}}_2 \\ \underset{\sim}{\mathbf{X}}_2 \end{pmatrix}$$

$$= \begin{pmatrix} I_q & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{p-q} \end{pmatrix} \begin{pmatrix} \underset{\sim}{X_1} \\ \underset{\sim}{X_2} \end{pmatrix}$$

$$= A\underset{\sim}{X} \qquad \rightarrow (1)$$

The joint p.d.f. of $\underset{\sim}{Y}$ is $g(\underset{\sim}{Y}) = g(\underset{\sim}{Y}_1)g(\underset{\sim}{Y}_2)$ ($\because \underset{\sim}{Y}_1$ & $\underset{\sim}{Y}_2$ are independent).

Also we know that (from the above theorem),

$$\underset{\sim}{Y}_1 \sim N_q(\underset{\sim}{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

$$\underset{\sim}{Y}_2 \sim N_{p-q}(\underset{\sim}{\mu}_2, \Sigma_{22})$$

$$\therefore g(\underset{\sim}{Y}) = n\left(\underset{\sim}{y}_{111} / \underset{\sim}{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).n\left(\underset{\sim}{y}_2 / \underset{\sim}{\mu}_2, \Sigma_{22}\right)$$

$$= \frac{1}{(2\pi)^{q/2}\left|\Sigma_{11.2}\right|}\exp\left\{-\frac{1}{2}(\underset{\sim}{Y}_1 - \underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2)'\Sigma_{11.2}^{-1}(\underset{\sim}{Y}_1 - \underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2)\right\}$$

$$= \frac{1}{(2\pi)^{(p-q)/2}\left|\Sigma_{22}\right|}\exp\left\{-\frac{1}{2}(\underset{\sim}{Y}_2 - \underset{\sim}{\mu}_2)'\Sigma_{22}^{-1}(\underset{\sim}{Y}_2 - \underset{\sim}{\mu}_2)\right\}$$

where , $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

If we make use of the linear transformation (non singular) as given in(1).The density function of $\underset{\sim}{X}$ is given by

$$f(\underset{\sim}{x}) = g(\underset{\sim}{y}(\underset{\sim}{x})).J(\underset{\sim}{x})$$

Where, $J(\underset{\sim}{X})$ is the jacobian and is given by

$$J(\underset{\sim}{X}) = \left|\frac{\partial \underset{\sim}{y}}{\partial \underset{\sim}{x}}\right| = |A| = |I_q|.|I_{p-q}| = 1$$

$$\therefore f(\underset{\sim}{X}) = f(\underset{\sim}{x}_1, \underset{\sim}{x}_2)$$

$$= \frac{1}{(2\pi)^{q/2}\left|\Sigma_{11.2}\right|} e^{-\frac{1}{2}(\underset{\sim}{x}_1 - \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{x}_2 - \underset{\sim}{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2)'\Sigma_{11.2}^{-1}(\underset{\sim}{x}_1 - \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{x}_2 - \underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2)}$$

$$\frac{1}{(2\pi)^{(p-q)/2}\left|\Sigma_{22}\right|} e^{-\frac{1}{2}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)'\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)} \qquad \text{-->(2)}$$

Now, By the definition conditional density of $\underset{\sim}{X}$, given that $\underset{\sim}{X}_2 = \underset{\sim}{x}_2$ is that

$$f(\underset{\sim}{x}_1 / \underset{\sim}{x}_2) = \frac{f(\underset{\sim}{x}_1, \underset{\sim}{x}_2)}{f(\underset{\sim}{x}_2)} \qquad \rightarrow (3)$$

where $f(\underset{\sim}{x}_1, \underset{\sim}{x}_2)$ is given by (2) and $f(\underset{\sim}{x}_2)$ is the marginal density of $\underset{\sim}{X}_2$ at the point $\underset{\sim}{x}_2$ where is nothing but $n(\underset{\sim}{x}_2/\underset{\sim}{\mu}_2, \Sigma_{22})$.

i.e. $f(\underset{\sim}{x}_2) = n(\underset{\sim}{x}_2/\underset{\sim}{\mu}_2, \Sigma_{22})$

$$= \frac{1}{(2\pi)^{(p-q)/2}\left|\Sigma_{22}\right|} \exp\left\{-\frac{1}{2}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)'\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)\right\}$$

Using (2) & (4) in (3) we get,

$$f(\underset{\sim}{x}_1/\underset{\sim}{x}_2) = \frac{1}{(2\pi)^{q/2}\left|\Sigma_{11.2}\right|} e^{-\frac{1}{2}[(\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1) - \Sigma_{12}\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)]'\Sigma_{11.2}^{-1}[(\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1) - \Sigma_{12}\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)]}$$

$\rightarrow$(5)

which is the conditional p.d.f. of $\underset{\sim}{X}_1$ given that $\underset{\sim}{X}_2 = \underset{\sim}{x}_2$.

From (5), it is clear that the density $f(\underset{\sim}{x}_1/\underset{\sim}{x}_2)$ is clearly a q-variate normal density with mean,

$$E(\underset{\sim}{X}_1/\underset{\sim}{X}_2 = \underset{\sim}{x}_2) = \underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)$$

$$= \nu(\underset{\sim}{x}_2) \quad \text{, say} \qquad \rightarrow (6)$$

and the variances matrix,

$$\text{var}(\underset{\sim}{x}_1/\underset{\sim}{X}_2 = \underset{\sim}{x}_2) = \Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad \rightarrow (7)$$

From (6)&(7) we may observe that the conditional mean of $\underset{\sim}{x}_1$ is simply a linear function of $\underset{\sim}{x}_2$ and the conditional co-variance of $\underset{\sim}{x}_1$ does not depend on $\underset{\sim}{x}_2$ at all.

The above result may be put in the following theorem:-

Let the components of $\underset{\sim}{X}$ be divided in to two groups composing the sub vectors $\underset{\sim}{X}_1$ & $\underset{\sim}{X}_2$. Suppose the mean $\underset{\sim}{\mu}$ is similarly divided into $\underset{\sim}{\mu}_1$ & $\underset{\sim}{\mu}_2$ and suppose the co-variance matrix Σ of $\underset{\sim}{X}$ is divided into $\Sigma_{11}, \Sigma_{12} = \Sigma_{21}, \Sigma_{22}$ the co-variance matrices of $\underset{\sim}{X}_1$ of $\underset{\sim}{X}_1$ & $\underset{\sim}{X}_2$, and of $\underset{\sim}{X}_2$ respectively. Then if the distribution of $\underset{\sim}{X}$ is normal, the conditional distribution of $\underset{\sim}{X}_1$ is given $\underset{\sim}{X}_2 = \underset{\sim}{x}_2$ is normal with mean $\underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)$ and co-variance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

NOTE:-

The above theorem may simply be asked as follows.If $X_1, X_2, ..., X_P$ have a joint normal distribution, then the conditional distribution of a subset of r.v's given that the remaining r.v's is also having normal distribution.

THE CHARACTERISTIC FUNCTION:-

DEFINITION:-

The characteristic function of a random vector $\underset{\sim}{\mathbf{X}}$ is

$$\phi(\underset{\sim}{t}) = E(e^{i\underset{\sim}{t}'\underset{\sim}{X}})$$

defined for every real vector $\underset{\sim}{t}$.

RESULT:-

If the components of a random vector $\underset{\sim}{\mathbf{X}}$ are independently distributed,

Then,

$$E(e^{i\underset{\sim}{t}'\underset{\sim}{X}}) = E\left(e^{i\sum_{j=1}^{p} t_j X_j}\right)$$

$$= \prod_{j=1}^{p} E(e^{it_j X_j})$$

THEOREM:-

The characteristic function of $\underset{\sim}{\mathbf{X}}$ which is distributed according to

$N(\underset{\sim}{\mu}, \Sigma)$ **is** $\quad \phi(\underset{\sim}{t}) = E(e^{i\underset{\sim}{t}'\underset{\sim}{X}}) = e^{i\underset{\sim}{t}'\underset{\sim}{\mu} - \frac{1}{2}\underset{\sim}{t}'\Sigma\underset{\sim}{t}}$

for every real vector $\underset{\sim}{t}$.

PROOF:-

We have given $\underset{\sim}{\mathbf{X}} \sim N(\underset{\sim}{\mu}, \Sigma)$.

Since, Σ and hence Σ^{-1} is symmetric and positive definite matrix there exists a non-singular matrix $\mathbf{C}'$ such that

$$\mathbf{C}'\Sigma^{-1}\mathbf{C} = \mathbf{I} \qquad \rightarrow (1)$$

$$\Rightarrow \Sigma^{-1} = (\mathbf{C}\mathbf{C}')^{-1} \text{ or } \Sigma = \mathbf{C}\mathbf{C}' \qquad \rightarrow (1.a)$$

we have the p.d.f. of $\underset{\sim}{\mathbf{X}}$ is

$$f(\underset{\sim}{\mathbf{X}}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})} \qquad \rightarrow (2)$$

Let us make use of the linear transformation,

$$\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}} = \mathbf{C}\underset{\sim}{\mathbf{Y}} \qquad (\mathbf{C} \text{ is defined as in (1)}) \qquad \rightarrow (3)$$

Then the p.d.f. of the new random vector $\underset{\sim}{\mathbf{Y}}$ is

$$g(\underset{\sim}{\mathbf{Y}}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}\underset{\sim}{\mathbf{y}}'\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}\underset{\sim}{\mathbf{y}}} \mathrm{J}(\underset{\sim}{\mathbf{y}}) \qquad \rightarrow (4)$$

where $\mathrm{J}(\underset{\sim}{\mathbf{y}})$ is the jacobian transformation and is given by

$$\mathrm{J}(\underset{\sim}{\mathbf{y}}) = \mathrm{mod}\left|\frac{\partial \underset{\sim}{\mathbf{x}}}{\partial \underset{\sim}{\mathbf{y}}}\right| = \mathrm{mod}|\mathbf{C}| \quad \left(\because \frac{\partial \underset{\sim}{\mathbf{x}}}{\partial \underset{\sim}{\mathbf{y}}} = \frac{\partial \mathbf{C}\underset{\sim}{\mathbf{y}} + \underset{\sim}{\boldsymbol{\mu}}}{\partial \underset{\sim}{\mathbf{y}}} = \mathbf{C}\right)$$

$$= \mathrm{mod}|\boldsymbol{\Sigma}|^{1/2} \qquad \left(\text{from 1(a)} |\boldsymbol{\Sigma}| = |\mathbf{C}\mathbf{C}'| = |\mathbf{C}|^2\right)$$

$$= |\boldsymbol{\Sigma}|^{1/2} \qquad \left(\because |\boldsymbol{\Sigma}| > 0\right)$$

Therefore (4) becomes from (1),

$$g(\underset{\sim}{\mathbf{Y}}) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}}} = n(\underset{\sim}{\mathbf{y}}/\underset{\sim}{\mathbf{0}}, I_p)$$

i.e. $\underset{\sim}{\mathbf{Y}} \sim N_p(\underset{\sim}{\mathbf{0}}, I_p)$.

The characteristic function of $\underset{\sim}{\mathbf{Y}}$ is

$$\phi(\underset{\sim}{\mathbf{u}}) = E(e^{\mathbf{i}\underset{\sim}{\mathbf{u}}'\underset{\sim}{\mathbf{y}}}) = E\left(e^{\mathbf{i}\sum_{j=1}^{p}\underset{\sim}{\mathbf{u}}_j'\mathbf{y}_j}\right)$$

$$= \prod_{j=1}^{p} E(e^{iu_j y_j}) \qquad (\because Y_j's \text{ are independent})$$

$$= \prod_{j=1}^{p} e^{-\frac{1}{2}u_j^2} \qquad (\because \text{the characteristic function of the standard normal variate } Y_j \text{ is } se^{-\frac{1}{2}u_j^2})$$

$$= e^{-\frac{1}{2}\sum_{j=1}^{p}u_j^2}$$

$$= e^{-\frac{1}{2}\underset{\sim}{\mathbf{u}}'\underset{\sim}{\mathbf{u}}} \qquad \rightarrow (5)$$

Now,

$$\phi(t) = E(e^{i\underset{\sim}{\mathbf{t}}'\underset{\sim}{\mathbf{X}}}) = E(e^{i\underset{\sim}{\mathbf{t}}'(\mathbf{C}\underset{\sim}{\mathbf{Y}} + \underset{\sim}{\boldsymbol{\mu}})}) \qquad \text{(from (3))}$$

i.e.

$$\phi(\underset{\sim}{t}) = E(e^{i\underset{\sim}{t}'\underset{\sim}{\mu}}) = e^{i\underset{\sim}{t}'\underset{\sim}{\mu}} E\left(e^{i\underset{\sim}{t}'\mathbf{C}\underset{\sim}{Y}}\right)$$

$$= e^{i\underset{\sim}{t}'\underset{\sim}{\mu}} E\left(e^{i\underset{\sim}{u}'\underset{\sim}{Y}}\right) \qquad (\text{where } \underset{\sim}{u}' = \underset{\sim}{t}'\mathbf{C})$$

$$= e^{i\underset{\sim}{t}'\underset{\sim}{\mu}} e^{-\frac{1}{2}\underset{\sim}{u}'\underset{\sim}{u}} \qquad (\text{from (5)})$$

$$= e^{i\underset{\sim}{t}'\underset{\sim}{\mu} - \frac{1}{2}\underset{\sim}{t}'\mathbf{C}\mathbf{C}'\underset{\sim}{t}} \qquad (\because \underset{\sim}{u}' = \underset{\sim}{t}'\mathbf{C})$$

$$= e^{i\underset{\sim}{t}'\underset{\sim}{\mu} - \frac{1}{2}\underset{\sim}{t}'\Sigma\underset{\sim}{t}} \qquad (\text{from 1.a})$$

Hence the proof .

THEOREM :-

If every linear combination of the components of a vector $\underset{\sim}{X}$ is normally distributed ,then $\underset{\sim}{X}$ is normally distributed **.**

PROOF:-

Suppose $\underset{\sim}{X}$ is a random vector of p random variables with mean vector $\underset{\sim}{\mu}$ and co-variance matrix $\mathbf{\Sigma}$.

Let us consider a linear combination of $\underset{\sim}{X}$ viz… $\underset{\sim}{c}'\underset{\sim}{X}$, where $\underset{\sim}{c}' = (c_1, c_2, ..., c_p)$.

We have given , $\underset{\sim}{c}'\underset{\sim}{X}$ is normal variate.

We have,

$$E(\underset{\sim}{c}'\underset{\sim}{X}) = \underset{\sim}{c}'E(\underset{\sim}{X})$$
$$= \underset{\sim}{c}'\underset{\sim}{\mu}$$
$$\mathrm{V}(\underset{\sim}{c}'\underset{\sim}{X}) = \mathrm{V}(\underset{\sim}{c}'\underset{\sim}{X})$$
$$= \underset{\sim}{c}'\mathrm{V}(\underset{\sim}{X})\underset{\sim}{c}$$
$$= \underset{\sim}{c}'\mathbf{\Sigma}\underset{\sim}{c} \qquad (\text{variance})$$

If may be noted $\underset{\sim}{c}'\underset{\sim}{\mu}$ **&** $\underset{\sim}{c}'\mathbf{\Sigma}\underset{\sim}{c}$ are scalars and they are respectively the mean & variances of the univariate random variables $\underset{\sim}{c}'\underset{\sim}{X}$. We have given that $\underset{\sim}{c}'\underset{\sim}{X} \sim N(\underset{\sim}{c}'\underset{\sim}{\mu}, \underset{\sim}{c}'\Sigma\underset{\sim}{c})$.

Let $Y = \underset{\sim}{c}'\underset{\sim}{X}$ and from the univariate normal distribution theory.

The characteristic function of Y is given by

$$\psi(t) = \mathrm{E}(e^{it\mathrm{Y}})$$

$$= e^{it\mathrm{E}(\mathrm{Y}) - \frac{1}{2}t^2\mathrm{V}(Y)}$$

$$= e^{it\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}} - \frac{1}{2}t^2\underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}}$$

If we write t=1 then, $\psi(t)$ becomes

$$\phi(\underset{\sim}{\mathbf{c}}) = e^{i\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}} - \frac{1}{2}\underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}} \qquad \text{where, } \underset{\sim}{\mathbf{X}} \sim N(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$$

which is the characteristic function of a multivariate random vector $\underset{\sim}{\mathbf{X}}$ whose mean vector is $\underset{\sim}{\boldsymbol{\mu}}$ & variance-covariance matrix is of $\boldsymbol{\Sigma}$.

But the mean & variance-covarince matrix of $\underset{\sim}{\mathbf{X}}$ respectively same as $\underset{\sim}{\boldsymbol{\mu}}$ & $\boldsymbol{\Sigma}$ and therefore, $\underset{\sim}{\mathbf{X}} \sim N(\underset{\sim}{\boldsymbol{\mu}}, \Sigma)$.

Hence the proof.

THOREM :-

If $\underset{\sim}{\mathbf{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \Sigma)$ then, $\underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}}$ is uni-normal variate with mean $\underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}}$ and variance $\underset{\sim}{\mathbf{c}}'\Sigma\underset{\sim}{\mathbf{c}}$.

(OR)

If $X_1, X_2, ..., X_p$ are jointly distributed as p-variate normal then its linear combination follows uni-variate normal distribution.

PROOF:-

$$\text{Let } \underset{\sim}{\mathbf{X}} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \Sigma$$

Then its characteristic function is given by

$$\phi(\underset{\sim}{\mathbf{t}}) = E(e^{i\underset{\sim}{\mathbf{t}}'\underset{\sim}{\mathbf{X}}}))$$

$$= e^{i\underset{\sim}{\mathbf{t}}'\underset{\sim}{\boldsymbol{\mu}} - \frac{1}{2}\underset{\sim}{\mathbf{t}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{t}}} \qquad \rightarrow (1)$$

$$\text{Let us write } \underset{\sim}{\mathbf{t}} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_p \end{pmatrix}$$

$$= \begin{pmatrix} t.c_1 \\ t.c_2 \\ \vdots \\ t.c_p \end{pmatrix}$$

$$= t\underset{\sim}{\mathbf{c}}$$

Then (1) becomes ,

$$\phi(\underset{\sim}{\mathbf{t}}) = e^{it\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}} - \frac{1}{2}t\underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}t}$$

$$= e^{it\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}} - \frac{1}{2}t^2\underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}}$$

$$= \mathrm{E}(e^{\mathrm{ity}})$$

$$= \psi(t) \quad \text{,say} \qquad \rightarrow (2)$$

where Y is normal variate with mean $\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}}$ and variance $\underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}$.

In other words (2) is the characteristic function $\psi(t)$ of a uni-normal variate whose mean is $\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}}$ and variance is $\underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}$.

If we consider the linear combination of the components of the normal random vector $\underset{\sim}{\mathbf{X}}$ viz.,

$$\mathrm{Y} = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}}$$

$$= c_1\mathrm{X}_1 + c_2\mathrm{X}_2 + \ldots + c_p\mathrm{X}_p$$

its mean and variance are given by

$$E(\mathrm{Y}) = E(\underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}}) = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}} \quad \& \quad \mathrm{V}(\mathrm{Y}) = \mathrm{V}(\underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}})$$

$$= \underset{\sim}{\mathbf{c}}'\mathrm{V}(\underset{\sim}{\mathbf{X}})\underset{\sim}{\mathbf{c}}$$

$$= \underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}$$

Thus, from the above explanation it follows that $\mathrm{Y} = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}}$ follows uvi-variate normal distribution,

i.e. , . $\mathrm{Y} = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}} \sim \mathrm{N}(\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}}, \underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}})$

Hence the proof.

SAMPLING FROM MULTINORMAL DISTRIBUTION

The Multivariate normal likelihood:

Let us assume that the $p \times 1$ vectors $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, \cdots, \underset{\sim}{\mathbf{X}}_n$ represent a random sample from a multivariate normal population with mean vector $\underset{\sim}{\boldsymbol{\mu}}$ and covariance matrix Σ. Since $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, \cdots, \underset{\sim}{\mathbf{X}}_n$ are mutually independent (by virtue of randomization) and each has distributed as the joint p.d.f. of all the observations is the product of the marginal normal densities.

i.e.

$$f(\underset{\sim}{x}_1, \underset{\sim}{x}_2 \ldots \underset{\sim}{x}_n) = f(\underset{\sim}{x}_1) \cdots f(\underset{\sim}{x}_n)$$

$$= \prod_{j=1}^{n} \left\{ \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}})} \right\}$$

$$\left[\because f(x_j) = n\left(\underset{\sim}{\mathbf{x}}_j / \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \right]$$

$$= \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n} (\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}})} \quad \text{--- (1)}$$

When the numerical value of the observations become available, there may be substituted for $\underset{\sim}{\mathbf{X}}_j$ in equation (1). The resulting expression, now considered as a function of $\underset{\sim}{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}$ and for a fixed set of observations $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \cdots, \underset{\sim}{x}_n$, it is called as "**the likelihood function**" and is denoted as $L(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$.

ML ESTIMATION OF MEAN VECTOR $\underset{\sim}{\mu}$ AND VARIANCE – COVARIANCE MATRIX Σ:

Consider the likelihood function $L(\underset{\sim}{\mu}, \Sigma)$ .given by (1). i.e.

$$L(\underset{\sim}{\mu}, \Sigma) = \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n} (\underset{\sim}{x}_j - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu})} \quad \text{---- (2)}$$

Now the maximum likelihood estimates of $\underset{\sim}{\mu}$ and Σ can be obtained by maximizing $L(\underset{\sim}{\mu}, \Sigma)$.

In order to obtain the MLE's of $\underset{\sim}{\mu}$ and Σ ,let us consider logarithms of (2) and is given by

$$\log L(\underset{\sim}{\mu}, \Sigma) = \frac{-np}{2} \log(2\pi) - \frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{j=1}^{n} (\underset{\sim}{x}_j - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu}) \quad \text{-->(3)}$$

-

Consider the last term of (3) and as if is a scalar we may write if as

$$\sum_{j=1}^{n} (\underset{\sim}{x}_j - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu})$$

$$= tr\left[\sum_{j=1}^{n} (\underset{\sim}{x}_j - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu}) \right]$$

$$= \sum_{j=1}^{n} tr\left[(\underset{\sim}{x}_j - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu}) \right]$$

$$= \sum_{j=1}^{n} tr\left[\Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu})(\underset{\sim}{x}_j - \underset{\sim}{\mu})' \right]$$

$$(\because \quad tr(AB) = tr(BA)$$

$$= tr\left[\boldsymbol{\Sigma}^{-1}\left\{\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\underset{\sim}{\boldsymbol{\mu}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\underset{\sim}{\boldsymbol{\mu}}\right)'\right\}\right] \qquad \text{-->(4)}$$

Now consider

$$\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\underset{\sim}{\boldsymbol{\mu}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\underset{\sim}{\boldsymbol{\mu}}\right)'$$

$$= \sum_{j=1}^{n}\left[\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}+\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right]\left[\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}+\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right]'$$

Where $\overline{\underset{\sim}{\mathbf{x}}} = \frac{1}{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{1}}+\underset{\sim}{\mathbf{x}}_{\mathbf{2}}+\cdots+\underset{\sim}{\mathbf{x}}_{\mathbf{n}}\right)$

$$= \sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)' + \sum_{j=1}^{n}\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right)\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right)'$$

(Since the cross product terms

$$\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right)' = \left(n\overline{\underset{\sim}{\mathbf{x}}}-n\overline{\underset{\sim}{\mathbf{x}}}\right)\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right)' = 0$$

$$\left[\because \sum_{j=1}^{n}\underset{\sim}{\mathbf{x}}_{\mathbf{j}} = n\,\overline{\underset{\sim}{\mathbf{x}}}\right]$$

and similarly

$$\sum_{j=1}^{n}\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)' = 0$$)

Thus

$$\sum_{j=1}^{n} \left(\underset{\sim}{x}_j - \underset{\sim}{\mu}\right)\left(\underset{\sim}{x}_j - \underset{\sim}{\mu}\right)'$$

$$= \sum_{j=1}^{n} \left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)' + n\left(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}\right)\left(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}\right)'$$

Substituting this in (4) we get ,

$$\sum_{j=1}^{n} \left(\underset{\sim}{x}_j - \underset{\sim}{\mu}\right)' \Sigma^{-1} \left(\underset{\sim}{x}_j - \underset{\sim}{\mu}\right)$$

$$= tr\left[\Sigma^{-1}\left\{\sum_{j=1}^{n}\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)' + n\left(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}\right)\left(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}\right)'\right\}\right]$$

$$= tr\left[\Sigma^{-1}\left\{\sum_{j=1}^{n}\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)'\right\}\right] + n\left[tr\left\{\left(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}\right)' \Sigma^{-1}\left(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}\right)\right\}\right]$$

[Since tr (AB) =tr (BA)]

$$= tr\left[\Sigma^{-1}\left\{\sum_{j=1}^{n}\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)'\right\}\right] + n\left\{\left(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}\right)' \Sigma^{-1}\left(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}\right)\right\}$$

--- (5)

(Since trace of scalar is scalar)

Substituting (5) in (3),we get

$$\log L\left(\underset{\sim}{\mu}, \Sigma\right) = \frac{-np}{2}\log(2\pi) + \frac{n}{2}\log\left|\Sigma^{-1}\right|$$

$$-\frac{1}{2}\left[tr\left\{ \boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)'\right)\right\}\right]$$

$$-\frac{n}{2}\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right)'\boldsymbol{\Sigma}^{-1}\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right) \quad \text{---- (6)}$$

Since $\boldsymbol{\Sigma}^{-1}$ is positive definite

$$\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right)'\boldsymbol{\Sigma}^{-1}\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right) > 0 \quad \forall\, \underset{\sim}{\boldsymbol{\mu}} \neq \overline{\underset{\sim}{\mathbf{x}}}$$

$$= 0 \quad \text{if } \underset{\sim}{\boldsymbol{\mu}} = \overline{\underset{\sim}{\mathbf{x}}}$$

From (6), we can observe that if the last term is zero then(6) becomes maximum that is $\log L\left(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$ can be maximized with respect to $\underset{\sim}{\boldsymbol{\mu}}$ at $\hat{\underset{\sim}{\boldsymbol{\mu}}} = \overline{\underset{\sim}{\mathbf{x}}}$

$\therefore$ The MLE of $\underset{\sim}{\boldsymbol{\mu}}$ is $\overline{\underset{\sim}{\mathbf{x}}}$ substituting the MLE of $\underset{\sim}{\boldsymbol{\mu}}$ ($\overline{\underset{\sim}{\mathbf{x}}}$) in (6)

We get

$$\log L\left(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right) = \frac{-np}{2}\log\left(2\pi\right) + \frac{n}{2}\log\left|\boldsymbol{\Sigma}^{-1}\right|$$

$$-\frac{1}{2}\left[tr\left\{ \boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)'\right)\right\}\right]$$

--- (7)

Now we have to maximize (7) w.r.t. $\boldsymbol{\Sigma}$ as the equation is free of $\underset{\sim}{\boldsymbol{\mu}}$

We can prove that (7) attains its maximum value at $\boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}}$,

Where, $$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)' \quad \text{----(8)}$$

Thus $\hat{\boldsymbol{\Sigma}}$ (given by (8)) is the MLE of $\boldsymbol{\Sigma}$.

The maximum value of the likelihood can be obtained by substituting the MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively given by

$$\hat{\boldsymbol{\mu}} = \underset{\sim}{\overline{\mathbf{x}}}$$

and $$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}} - \underset{\sim}{\overline{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}} - \underset{\sim}{\overline{\mathbf{x}}}\right)'$$

in (2) and it is given by

$$L\left(\underset{\sim}{\hat{\mu}}, \hat{\Sigma}\right) = (2\pi)^{-np/2}\left|\hat{\boldsymbol{\Sigma}}\right|^{-n/2} e^{\frac{-1}{2}tr\left[n\left(\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}} - \underset{\sim}{\overline{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}} - \underset{\sim}{\overline{\mathbf{x}}}\right)'\right)^{-1}\left(\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}} - \underset{\sim}{\overline{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}} - \underset{\sim}{\overline{\mathbf{x}}}\right)'\right)\right]}$$

$$= (2\pi)^{-np/2}\left|\hat{\boldsymbol{\Sigma}}\right|^{-n/2} e^{\frac{-n}{2}tr\left(\mathrm{I}_{\mathrm{p}}\right)}$$

$$= (2\pi)^{-np/2}\left|\hat{\boldsymbol{\Sigma}}\right|^{-n/2} e^{-np/2}$$

$$= \text{const. x } \left|\hat{\boldsymbol{\Sigma}}\right|^{-\frac{n}{2}}$$

$$= \text{const. x } (\text{generilised variance})^{\frac{-n}{2}}$$

since generalized variance is defined as $\left|\hat{\boldsymbol{\Sigma}}\right|$ .The generalized variance determines the *peakedness* of the likelihood function and consequently is a natural measure of variability when the parent population is multivariate normal.

NOTE:-

1.MLE's posses an invariance property which means if $\hat{\theta}$ is the MLE of θ then $h(\hat{\theta})$ is the MLE of $h(\theta)$,where $h(\theta)$ is a function of θ .

For Example :-

i.If $\hat{\underset{\sim}{\mu}}$ is MLE of $\hat{\underset{\sim}{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ is MLE of $\boldsymbol{\Sigma}$ then $\hat{\underset{\sim}{\mu}}^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\underset{\sim}{\mu}}$ is MLE of $\underset{\sim}{\mu}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}$.

ii.If σ_{ij} is the ij^{th} element of $\boldsymbol{\Sigma}$ and $\hat{\sigma}_{ij}$ is the ij^{th} element of $\hat{\boldsymbol{\Sigma}}$ where $\hat{\boldsymbol{\Sigma}}$ is the MLE of σ_{ij} .

Where $$\hat{\sigma}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\left(X_{ik} - \bar{X}_i\right)\left(X_{jk} - \bar{X}_j\right)$$

$$= COV\left(X_i, X_j\right).$$

2.From equation (6) the log-likelihood and hence the joint p.d..f depends on the whole set of observations $\underset{\sim}{x}_1 , \cdots , \underset{\sim}{x}_n$ only through the sample mean $\bar{\underset{\sim}{\mathbf{x}}}$ and the sum of squares and cross product matrix,

$$\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_j - \bar{\underset{\sim}{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_j - \bar{\underset{\sim}{\mathbf{x}}}\right)' = n\hat{\boldsymbol{\Sigma}}$$

We may express this fact by saying that $\hat{\underset{\sim}{\mu}}$ $\left(\text{or } \bar{\underset{\sim}{\mathbf{x}}}\right)$ and $\hat{\boldsymbol{\Sigma}}$ are sufficient statistics . Thus the MLEs $\hat{\underset{\sim}{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are sufficient statistics of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

3.The MLE of $\boldsymbol{\Sigma}$ is

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)'$$

Thos formula is not convenient to compute $\hat{\Sigma}$ and the following is the convenient formula for computation

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{x}_j\underset{\sim}{x}_j' - \bar{\underset{\sim}{x}}\,\bar{\underset{\sim}{x}}'$$

Explanation:-

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)'$$

$$= \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{x}_j\underset{\sim}{x}_j' - \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{x}_j\bar{\underset{\sim}{x}}' - \frac{1}{n}\sum_{j=1}^{n}\bar{\underset{\sim}{x}}\underset{\sim}{x}_j' + \frac{1}{n}\sum_{j=1}^{n}\bar{\underset{\sim}{x}}\,\bar{\underset{\sim}{x}}'$$

$$= \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{x}_j\underset{\sim}{x}_j' - \bar{\underset{\sim}{x}}\,\bar{\underset{\sim}{x}}' - \bar{\underset{\sim}{x}}\bar{\underset{\sim}{x}}' + \bar{\underset{\sim}{x}}\bar{\underset{\sim}{x}}'$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{x}_j\underset{\sim}{x}_j' - \bar{\underset{\sim}{x}}\,\bar{\underset{\sim}{x}}'$$

Sampling Distribution of the MLE's $\hat{\underset{\sim}{\mu}}$ and $\hat{\Sigma}$ and the independence of $\hat{\underset{\sim}{\mu}}$ and $\hat{\Sigma}$:- Before going to obtain the sampling distribution of $\hat{\underset{\sim}{\mu}}$ and $\hat{\Sigma}$,let us prove

the following result which is useful in obtaining the sampling distributions of $\hat{\underset{\sim}{\mu}}$ and $\hat{\Sigma}$.

Result:-

Suppose $\underset{\sim}{X}_1, \cdots, \underset{\sim}{X}_n$ are independent where $\underset{\sim}{X}_\alpha \sim Np\left(\underset{\sim}{\mu}_\alpha, \Sigma\right)$.

Let

$C = \left(c_{\alpha j}\right)_{n \times n}$ be an orthogonal matrix then

$$\underset{\sim}{Y}_\alpha = \sum_{j=1}^{n} c_{\alpha j} \underset{\sim}{X}_j \sim Np\left(\underset{\sim}{\upsilon}_\alpha, \Sigma\right)$$

Where $\underset{\sim}{\upsilon}_\alpha = \sum_{j=1}^{n} c_{\alpha j} \underset{\sim}{\mu}_j$ and

$\underset{\sim}{Y}_1, \cdots, \underset{\sim}{Y}_n$ are independent.

Proof : - Since 'C' is orthogonal matrix ,

We have $\sum_{j=1}^{n} c_{\alpha j}^2 = 1$ and $\sum_{j=1}^{n} c_{\alpha j} c_{\beta j} = 0$ ---- (1)

In order to prove $\underset{\sim}{Y}_\alpha \sim Np\left(\underset{\sim}{\upsilon}_\alpha, \Sigma\right)$, Let us consider the characteristic function of $\underset{\sim}{Y}_\alpha$ is

$$E\left(e^{i \underset{\sim}{t}' \underset{\sim}{Y}_\alpha}\right) = E\left(e^{i \underset{\sim}{t}' \sum_{j=1}^{n} C_{\alpha j} \underset{\sim}{X}_j}\right)$$

$$= E\left(\prod_{j=1}^{n} e^{iC_{\alpha j}\underset{\sim}{t}'\underset{\sim}{X}_j}\right)$$

$$= \prod_{j=1}^{n} E\left(e^{i\underset{\sim}{u}'\underset{\sim}{X}_j}\right)$$

(Since $\underset{\sim}{X}_1, \cdots, \underset{\sim}{X}_n$ are independent and where $\underset{\sim}{u} = C_{\alpha j}\underset{\sim}{t}$)

$$= \prod_{j=1}^{n} e^{i\underset{\sim}{u}'\underset{\sim}{\mu}_j - \frac{1}{2}\underset{\sim}{u}'\Sigma\,\underset{\sim}{u}}$$

$$\left(\because \underset{\sim}{X}_j \sim Np\left(\underset{\sim}{\mu}_j, \Sigma\right)\right)$$

$$= e^{i\sum_{j=1}^{n}\underset{\sim}{u}'\underset{\sim}{\mu}_j - \frac{1}{2}\sum_{j=1}^{n}\underset{\sim}{u}'\Sigma\underset{\sim}{u}}$$

$$= e^{i\underset{\sim}{t}'\sum_{j=1}^{n} C_{\alpha j}\underset{\sim}{\mu}_j - \frac{1}{2}\underset{\sim}{t}'\Sigma\underset{\sim}{t}\sum_{j=1}^{n} C_{\alpha j}^2}$$

(Substituting $\underset{\sim}{u} = C_{\alpha j}\underset{\sim}{t}$)

$$= e^{i\underset{\sim}{t}'\sum_{j=1}^{n} C_{\alpha j}\underset{\sim}{\mu}_j - \frac{1}{2}\underset{\sim}{t}'\Sigma\,\underset{\sim}{t}}$$

[Using (1)]

Which is the characteristic function of a multivariate normal vector whose mean vector

$\sum_{j=1}^{n} C_{\alpha j} \underset{\sim}{\mu}_{j}$ & variance-covariance matrix $\sum$.

Thus $\underset{\sim}{Y}_{\alpha} \sim Np\left(\underset{\sim}{\upsilon}_{\alpha}, \Sigma\right)$, Where $\underset{\sim}{\upsilon}_{\alpha} = \sum_{j=1}^{n} C_{\alpha j} \underset{\sim}{\mu}_{j}$

Now if remains to prove that $\underset{\sim}{Y}_{1}, \cdots, \underset{\sim}{Y}_{n}$ are independent.

In order to prove $\underset{\sim}{Y}_{1}, \ldots, \underset{\sim}{Y}_{n}$ are independent we have to prove that (as $\underset{\sim}{Y}_{j}$'s are multivariate normal vector).

$$\operatorname{cov}\left(\underset{\sim}{Y}_{\alpha}, \underset{\sim}{Y}_{\beta}\right) = O_{p \times p} \quad \text{(Zero matrix)}$$

The covariance matrix between $\underset{\sim}{Y}_{\alpha}$ and $\underset{\sim}{Y}_{\beta}$ is

$$\operatorname{cov}\left(\underset{\sim}{Y}_{\alpha}, \underset{\sim}{Y}_{\beta}\right) = E\left[\left(\underset{\sim}{Y}_{\alpha} - E\left(\underset{\sim}{Y}_{\alpha}\right)\right)\left(\underset{\sim}{Y}_{\beta} - E\left(\underset{\sim}{Y}_{\beta}\right)\right)'\right]$$

$$= E\left[\left(\underset{\sim}{Y}_{\alpha} - \underset{\sim}{\upsilon}_{\alpha}\right)\left(\underset{\sim}{Y}_{\beta} - \underset{\sim}{\upsilon}_{\beta}\right)'\right]$$

$$= E\left[\left(\sum_{i=1}^{n} C_{\alpha i} \underset{\sim}{X}_{i} - \sum_{i=1}^{n} C_{\alpha i} \underset{\sim}{\mu}_{i}\right)\left(\sum_{j=1}^{n} C_{\beta j} \underset{\sim}{X}_{j} - \sum_{j=1}^{n} C_{\beta j} \underset{\sim}{\mu}_{j}\right)'\right]$$

$$= E\left[\left(\sum_{i=1}^{n} C_{\alpha i}\left(\underset{\sim}{\mathbf{X}}_i - \underset{\sim}{\boldsymbol{\mu}}_i\right)\sum_{j=1}^{n} C_{\beta j}\left(\underset{\sim}{\mathbf{X}}_j - \underset{\sim}{\boldsymbol{\mu}}_j\right)\right)'\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} C_{\alpha i}C_{\beta j}E\left[\left(\underset{\sim}{\mathbf{X}}_i - \underset{\sim}{\boldsymbol{\mu}}_i\right)\left(\underset{\sim}{\mathbf{X}}_j - \underset{\sim}{\boldsymbol{\mu}}_j\right)'\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} C_{\alpha i}C_{\beta j}\,\mathrm{cov}\left(\underset{\sim}{\mathbf{X}}_i, \underset{\sim}{\mathbf{X}}_j'\right)$$

$$= \sum_{i=1}^{n} c_{\alpha i}c_{\beta i}V(\underset{\sim}{X}_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} c_{\alpha i}c_{\beta j}\,\mathrm{cov}(\underset{\sim}{X}_i, \underset{\sim}{X}_j)$$

since $\underset{\sim}{X}_i'$ are independent $\mathrm{cov}(\underset{\sim}{X}_i, \underset{\sim}{X}_j) = 0$

There fore the second term of the above will vanish.

There fore now

$$\mathrm{cov}(\underset{\sim}{Y}_\alpha, \underset{\sim}{Y}_\beta') = \sum_{i=1}^{n} c_{\alpha i}c_{\beta j}V(\underset{\sim}{X}_i)$$

$$= \sum\sum_{i=1}^{n} c_{\alpha i}c_{\beta i}$$

$$(\because V(\underset{\sim}{X}_i) = \textstyle\sum)$$

But from equation (1)

$$\sum_{i=1}^{n} c_{\alpha i}c_{\beta i} = 0$$

$$\therefore \mathrm{cov}(\underset{\sim}{Y}_\alpha, \underset{\sim}{Y}_\beta') = 0$$

Thus $\underset{\sim}{\mathbf{Y}}_{\boldsymbol{\alpha}}$ and $\underset{\sim}{\mathbf{Y}}_{\boldsymbol{\beta}}$ are independent and consequently $\underset{\sim}{\mathbf{Y}}_1, \cdots, \underset{\sim}{\mathbf{Y}}_n$ are independent. Hence the theorem.

Theorem:-

Let $\underset{\sim}{\mathbf{X}}_1, \cdots, \underset{\sim}{\mathbf{X}}_n$ be an independent random sample from $N_p\left(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$.

Then the MLE of $\boldsymbol{\mu}$ say $\hat{\boldsymbol{\mu}}$ (also the sample mean) is distributed according to $N_p\left(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}/n\right)$ and is independent of the MLE of $\boldsymbol{\Sigma}$ given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{\alpha=1}^{n}\left(\underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}} - \underset{\sim}{\bar{\mathbf{X}}}\right)\left(\underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}} - \underset{\sim}{\bar{\mathbf{X}}}\right)'$$

and $n\hat{\boldsymbol{\Sigma}}$ is distributed as $\sum_{\alpha=1}^{n}\underset{\sim}{\mathbf{Z}}_{\boldsymbol{\alpha}}\underset{\sim}{\mathbf{Z}}'_{\boldsymbol{\alpha}}$ where $\underset{\sim}{\mathbf{Z}}_{\boldsymbol{\alpha}} \sim N_p\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\right)$ and is $\underset{\sim}{Z}_1, \ldots\ldots \underset{\sim}{Z}_n$ are independent.

Proof:-

We have given a random sample $\underset{\sim}{\mathbf{X}}_1, \cdots, \underset{\sim}{\mathbf{X}}_n$ where $\underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}} \sim N_p\left(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$ and is independent of $\underset{\sim}{\mathbf{X}}_{\boldsymbol{\beta}}$ for $\alpha \neq \beta$ .We have the MLE's of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively given by

$$\hat{\mu} = 1/n\sum_{\alpha=1}^{n}\underset{\sim}{X}_{\alpha} = \underset{\sim}{\bar{X}}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{\alpha=1}^{n}\left(\underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}} - \underset{\sim}{\bar{\mathbf{X}}}\right)\left(\underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}} - \underset{\sim}{\bar{\mathbf{X}}}\right)' \quad \text{---}$$

(1)

Now there exists an $n\,x\,n$ orthogonal matrix $\mathbf{B} = \left(b_{\alpha\beta}\right)$ with the last row i.e.

$$b_{n\beta} = \frac{1}{\sqrt{n}} \forall \beta$$

$$\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}} \right) \qquad \text{--- (1.a)}$$

Let us define a new random sample $\underset{\sim}{\mathbf{Z}}_1, \cdots, \underset{\sim}{\mathbf{Z}}_n$ from the given random sample $\underset{\sim}{\mathbf{X}}_1, \cdots, \underset{\sim}{\mathbf{X}}_n$ using the orthogonal transformation from the orthogonal matrix B. Thus

$$\underset{\sim}{\mathbf{Z}}_\alpha = \sum_{\beta=1}^{n} b_{\alpha\beta} \underset{\sim}{\mathbf{X}}_\beta \qquad \text{for } \alpha = 1, 2, \cdots, n \qquad \text{---}$$

(2)

In particular,

$$\underset{\sim}{\mathbf{Z}}_n = \sum_{\beta} b_{n\beta} \underset{\sim}{\mathbf{X}}_\beta$$

$$= \sum_{\beta} \frac{1}{\sqrt{n}} \underset{\sim}{\mathbf{X}}_\beta \qquad \text{[The last row of } \mathrm{B} \text{ is as given in (1.a)]}$$

$$= \sqrt{n} \underset{\sim}{\bar{\mathbf{X}}} \qquad \text{[From (1)]} \qquad \text{--- (3)}$$

Let us consider

$$\sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{Z}}_\alpha \underset{\sim}{\mathbf{Z}}'_\alpha = \sum_{\alpha=1}^{n} \left(\sum_{i=1}^{n} b_{\alpha i} \underset{\sim}{\mathbf{X}}_i \right) \left(\sum_{j=1}^{n} b_{\alpha j} \underset{\sim}{\mathbf{X}}_j \right)' \qquad \text{[Using (2)]}$$

i.e.

$$= \sum_{\alpha=1}^{n} \sum_{i=1}^{n} \sum_{j=1}^{n} b_{\alpha i} b_{\alpha j} \underset{\sim}{X}_{i} \underset{\sim}{X}_{j}'$$

$$= \sum_{i=j=1}^{n} \sum_{\alpha=1}^{n} b_{\alpha i}^{2} \underset{\sim}{\mathbf{X}}_{\mathbf{i}} \underset{\sim}{\mathbf{X}}_{\mathbf{i}}' + \sum_{i \neq j=1}^{n} \sum_{\alpha=1}^{n} b_{\alpha i} b_{\alpha j} \underset{\sim}{\mathbf{X}}_{\mathbf{i}} \underset{\sim}{\mathbf{X}}_{\mathbf{j}}'$$

$$= \sum_{i=1}^{n} \underset{\sim}{\mathbf{X}}_{\mathbf{i}} \underset{\sim}{\mathbf{X}}_{\mathbf{i}}' \left(\sum_{\alpha=1}^{n} b_{\alpha i}^{2} \right) + \sum_{i \neq j=1}^{n} \underset{\sim}{\mathbf{X}}_{\mathbf{i}} \underset{\sim}{\mathbf{X}}_{j}' \left(\sum_{\alpha=1}^{n} b_{\alpha i} b_{\alpha j} \right)$$

$$= \sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}} \underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}}' \qquad \text{--- (4)}$$

[$\because$ B is the orthogonal matrix and as a consequence

$$\sum_{\alpha=1}^{n} b_{\alpha i} b_{\beta j} = 0 \quad \text{and} \quad \sum_{\alpha=1}^{n} b_{\alpha i}^{2} = 1 \quad]$$

Now consider $n\hat{\boldsymbol{\Sigma}}$ from (1) i.e.

$$n\hat{\boldsymbol{\Sigma}} = \sum_{\alpha=1}^{n} \left(\underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}} - \overline{\underset{\sim}{\mathbf{X}}} \right) \left(\underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}} - \overline{\underset{\sim}{\mathbf{X}}} \right)'$$

$$= \sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}} \underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}}' - \sum_{\alpha=1}^{n} \overline{\underset{\sim}{\mathbf{X}}} \overline{\underset{\sim}{\mathbf{X}}}'$$

$$\left(\because \overline{\mathbf{X}} \sum_{\alpha=1}^{n} \left(\underset{\sim}{\mathbf{X}}_{\boldsymbol{\alpha}} - \overline{\underset{\sim}{\mathbf{X}}} \right)' = 0 \right)$$

$$= \sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{X}}_{\alpha} \underset{\sim}{\mathbf{X}}'_{\alpha} - n\bar{\underset{\sim}{\mathbf{X}}}\bar{\underset{\sim}{\mathbf{X}}}'$$

$$= \sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha} - \underset{\sim}{\mathbf{Z}}_{\mathbf{n}} \underset{\sim}{\mathbf{Z}}'_{\mathbf{n}} \qquad \text{[Using (3) and (4)]}$$

$$= \sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha} \qquad \textbf{--- (5)}$$

From (3) and (5) we observe that $\bar{\underset{\sim}{\mathbf{X}}}$ ($\hat{\underset{\sim}{\boldsymbol{\mu}}}$) is distributed according to the distribution of $\underset{\sim}{\mathbf{Z}}_{\mathbf{n}}$ and $n\hat{\boldsymbol{\Sigma}}$ (and hence $\hat{\boldsymbol{\Sigma}}$)is distributed according to the distribution of $\underset{\sim}{\mathbf{Z}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{Z}}_{\mathbf{n-1}}$.

Also, since $\underset{\sim}{\mathbf{Z}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{Z}}_{\mathbf{n}}$ are obtained from $\underset{\sim}{\mathbf{X}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{X}}_{\mathbf{n}}$ using the orthogonal linear transformation (using orthogonal matrix $\mathbf{B}$) $\underset{\sim}{\mathbf{Z}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{Z}}_{\mathbf{n}}$ are independently distributed as Multivariate normal distribution with common covariance matrix ' $\boldsymbol{\Sigma}$ '.Therefore $\hat{\underset{\sim}{\boldsymbol{\mu}}}$ and $\hat{\boldsymbol{\Sigma}}$ are independently distributed.

Now let us obtain the mean vector of $\underset{\sim}{\mathbf{Z}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{Z}}_{\mathbf{n}}$

From (3)

$$E\left(\underset{\sim}{\mathbf{Z}}_{\mathbf{n}}\right) = \sqrt{n}\,E\left(\bar{\underset{\sim}{\mathbf{X}}}\right)$$

$$= \sqrt{n}\,\frac{1}{n}\left(E\left(\underset{\sim}{\mathbf{X}}_{\mathbf{1}} + \underset{\sim}{\mathbf{X}}_{\mathbf{2}} + \cdots + \underset{\sim}{\mathbf{X}}_{\mathbf{n}}\right)\right)$$

[$\because \underset{\sim}{\mathbf{X}}_{\mathbf{i}}{}'s$ are independent]

$$= \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) \quad = \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{\mu} = \sqrt{n}\underset{\sim}{\mu}$$

$$\left(\because \underset{\sim}{\mathbf{X_i}} \sim Np\left(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)\right)$$

Thus $\underset{\sim}{\mathbf{z_n}} \sim Np\left(\sqrt{n}\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$

i.e. $\sqrt{n}\,\underset{\sim}{\bar{\mathbf{X}}} \sim Np\left(\sqrt{n}\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$

i.e. $\underset{\sim}{\bar{\mathbf{X}}} \sim Np\left(\underset{\sim}{\boldsymbol{\mu}}, \frac{\boldsymbol{\Sigma}}{n}\right)$

From (2), we have

$$E\left(\underset{\sim}{\mathbf{z}_\alpha}\right) = \sum_{\beta=1}^{n} b_{\alpha\beta} E\left(\underset{\sim}{\mathbf{X}_\beta}\right) \quad [\ \because \underset{\sim}{\mathbf{X}_\beta}\text{'}s \text{ are independent }]$$

$$= \sum_{\beta=1}^{n} b_{\alpha\beta}\,\underset{\sim}{\boldsymbol{\mu}}$$

$$= \sum_{\beta=1}^{n} b_{\alpha\beta}\,\frac{1}{\sqrt{n}}\sqrt{n}\,\underset{\sim}{\boldsymbol{\mu}}$$

$$= \sqrt{n}\,\underset{\sim}{\boldsymbol{\mu}} \sum_{\beta=1}^{n} b_{\alpha\beta} b_{n\beta} \qquad [\ \because b_{n\beta} = \frac{1}{\sqrt{n}}\ ]$$

$$= \underset{\sim}{\mathbf{0}} \quad \forall\ \alpha \neq n$$

Thus each of $\underset{\sim}{\mathbf{Z}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{Z}}_{\mathbf{n-1}}$ are distributed as $Np\left(\underset{\sim}{0}, \boldsymbol{\Sigma}\right)$ .Therefore from (5) $n\hat{\boldsymbol{\Sigma}}$ is distributed as $\sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\boldsymbol{\alpha}} \underset{\sim}{\mathbf{Z}}'_{\boldsymbol{\alpha}}$, where $\underset{\sim}{\mathbf{Z}}_{\boldsymbol{\alpha}} \sim N\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\right)$ and is independent of $\underset{\sim}{\mathbf{Z}}_{\boldsymbol{\beta}} \left(\beta \neq \alpha\right)$

Thus the MLE's of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are independently distributed.

Hence the proof.

NOTE:-

- Since $\underset{\sim}{\mathbf{X}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{X}}_{\mathbf{n}}$ is a random sample

$$E\left(\underset{\sim}{\overline{\mathbf{X}}}\right) = \frac{1}{n}\left(E\left(\underset{\sim}{\mathbf{X}}_{\mathbf{1}}\right) + E\left(\underset{\sim}{\mathbf{X}}_{\mathbf{2}}\right) + \cdots + E\left(\underset{\sim}{\mathbf{X}}_{\mathbf{n}}\right)\right)$$

$$= \frac{n\,\underset{\sim}{\boldsymbol{\mu}}}{n} = \underset{\sim}{\boldsymbol{\mu}}$$

Thus $\underset{\sim}{\overline{\mathbf{X}}}$ is an unbiased estimator of $\underset{\sim}{\boldsymbol{\mu}}$ .Thus sample mean is an unbiased estimator of the population mean vector $\underset{\sim}{\boldsymbol{\mu}}$.

- $$E\left(\hat{\boldsymbol{\Sigma}}\right) = \frac{1}{n} E\left(\sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\boldsymbol{\alpha}} \underset{\sim}{\mathbf{Z}}'_{\boldsymbol{\alpha}}\right)$$

$$= \frac{1}{n} \sum_{\alpha=1}^{n-1} E\left(\underset{\sim}{\mathbf{Z}}_{\boldsymbol{\alpha}} \underset{\sim}{\mathbf{Z}}'_{\boldsymbol{\alpha}}\right) \quad [\because \underset{\sim}{Z}_{\alpha}{}'s \text{ are independent}]$$

$$= \frac{1}{n}\sum_{\alpha=1}^{n-1} V\left(\underset{\sim}{\mathbf{z}}_{\alpha}\right) \quad [\because E\left(\underset{\sim}{\mathbf{z}}_{\alpha} = \underset{\sim}{\mathbf{0}}\right)]$$

$$= \frac{1}{n}\sum_{\alpha=1}^{n-1} \boldsymbol{\Sigma} = \frac{n-1}{n}\boldsymbol{\Sigma}$$

Thus $\hat{\boldsymbol{\Sigma}}$ is not an unbiased estimator of $\boldsymbol{\Sigma}$. But

$$\frac{n}{n-1}\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{\alpha=1}^{n}\left(\underset{\sim}{\mathbf{X}}_{\alpha} - \overline{\underset{\sim}{\mathbf{X}}}\right)\left(\underset{\sim}{\mathbf{X}}_{\alpha} - \overline{\underset{\sim}{\mathbf{X}}}\right)' = \mathbf{S}$$

(say) is an unbiased estimator of $\boldsymbol{\Sigma}$ $\left[\because E\left(\frac{n}{n-1}\hat{\boldsymbol{\Sigma}}\right) = \boldsymbol{\Sigma}\right]$

Hence $\mathbf{S} = \frac{1}{n-1}\sum_{\alpha=1}^{n}\left(\underset{\sim}{\mathbf{X}}_{\alpha} - \overline{\underset{\sim}{\mathbf{X}}}\right)\left(\underset{\sim}{\mathbf{X}}_{\alpha} - \overline{\underset{\sim}{\mathbf{X}}}\right)'$ is called the sample covariance matrix and is an unbiased estimator of $\boldsymbol{\Sigma}$.

- From the above theorem, it is obviously follows that the sample mean ($\overline{\underset{\sim}{\mathbf{X}}}$) and the sample covariance matrix

$$\left(\mathbf{S} = \frac{1}{n-1}\sum_{\alpha=1}^{n}\left(\underset{\sim}{\mathbf{X}}_{\alpha} - \overline{\underset{\sim}{\mathbf{X}}}\right)\left(\underset{\sim}{\mathbf{X}}_{\alpha} - \overline{\underset{\sim}{\mathbf{X}}}\right)'\right)$$

are independently distributed. Also it may be seen that

$$\overline{\underset{\sim}{\mathbf{X}}} \sim N\left(\underset{\sim}{\boldsymbol{\mu}}, \frac{\boldsymbol{\Sigma}}{n}\right)$$

$$S = \frac{1}{n-1}\sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\alpha}\underset{\sim}{\mathbf{Z}}'_{\alpha} \quad \text{, where } \underset{\sim}{\mathbf{Z}}_{\alpha} \sim N_p\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\right)$$

i.e. **(n-1)S** is distributed according to the distribution of $\sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\alpha}\underset{\sim}{\mathbf{Z}}'_{\alpha}$, where $\underset{\sim}{\mathbf{Z}}_{1}, \cdots, \underset{\sim}{\mathbf{Z}}_{n}$ are independently distributed as $N_p\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\right)$ .The matrix $\sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\alpha}\underset{\sim}{\mathbf{Z}}'_{\alpha}$ is called " Wishart random matrix" and it is distributed according to "wishart distribution" with (n-1) degrees of freedom.

And is denoted as $W_{n-1}\left(\boldsymbol{\Sigma}\right)$, where $\boldsymbol{\Sigma}$ is the covariance matrix of Wishart random matrix .

Hence it may be noted that

$$\boldsymbol{\Sigma} = E\left(\frac{1}{n-1}\sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\alpha}\underset{\sim}{\mathbf{Z}}'_{\alpha}\right)$$

$$= E\left(\frac{\text{Wishart random matrix}}{\text{degrees of freedom}}\right)$$

Thus **(n-1)S** (and hence **S**) provides independent information about $\boldsymbol{\Sigma}$ and the distribution os S does not depend on $\underset{\sim}{\boldsymbol{\mu}}$ .This allows us to construct a statistics for making inferences about $\underset{\sim}{\boldsymbol{\mu}}$ as we shall see in the later section.

Properties of the Wishart distribution:-

- If $\mathbf{A}_1$ is distributed as $\mathrm{W}_{m_1}\left(\boldsymbol{\Sigma}\right)$ independently of $\mathbf{A}_2$, which is distributed as $\mathrm{W}_{m_2}\left(\boldsymbol{\Sigma}\right)$, then $\mathbf{A}_1+\mathbf{A}_2$ is distributed as $\mathrm{W}_{m_1+m_2}\left(\boldsymbol{\Sigma}\right)$. That is the degrees of freedom are added.

Proof: - Since $\mathbf{A}_1 \sim \mathrm{W}_{m_1}\left(\boldsymbol{\Sigma}\right)$.

$\mathbf{A}_1$ may be written as

$$\mathbf{A}_1=\sum_{\alpha=1}^{m_1}\underset{\sim}{\mathbf{z}}_{\boldsymbol{\alpha}}\underset{\sim}{\mathbf{z}}'_{\boldsymbol{\alpha}} \quad \text{, where} \quad \underset{\sim}{\mathbf{z}}_{\boldsymbol{\alpha}} \sim Np\left(\underset{\sim}{\mathbf{0}},\boldsymbol{\Sigma}\right)$$

Also since $\mathbf{A}_2$ is independently distributed as $\mathrm{W}_{m_2}\left(\boldsymbol{\Sigma}\right)$,
We may write

$$\mathbf{A}_2=\sum_{\alpha=m_1+1}^{m_1+m_2}\underset{\sim}{\mathbf{z}}_{\boldsymbol{\alpha}}\underset{\sim}{\mathbf{z}}'_{\boldsymbol{\alpha}} \quad \text{, where} \quad \underset{\sim}{\mathbf{z}}_{\boldsymbol{\alpha}} \sim Np\left(\underset{\sim}{\mathbf{0}},\boldsymbol{\Sigma}\right)$$

Since $\mathbf{A}_1$ and $\mathbf{A}_2$ are independent, $\underset{\sim}{\mathbf{z}}_{\mathbf{1}},\ldots,\underset{\sim}{\mathbf{z}}_{\mathbf{m_1+m_2}}$ are independent
and as a consequence .

$$\mathbf{A}=\mathbf{A}_1+\mathbf{A}_2=\sum_{\alpha=1}^{m_1+m_2}\underset{\sim}{\mathbf{z}}_{\boldsymbol{\alpha}}\underset{\sim}{\mathbf{z}}'_{\boldsymbol{\alpha}} \sim \mathrm{W}_{m_1+m_2}\left(\boldsymbol{\Sigma}\right)$$

Hence the proof.

- If $\mathbf{A} \sim \mathrm{W}_m\left(\boldsymbol{\Sigma}\right)$,then $\mathbf{CAC}' \sim \mathrm{W}_m\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)$

Proof:-Given $\mathbf{A} \sim \mathrm{W}_m\left(\boldsymbol{\Sigma}\right)$

$$\therefore \ \mathbf{A} = \sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha} \quad \text{, where } \underset{\sim}{\mathbf{Z}}_{\alpha} \sim N_p\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\right)$$

$$\mathbf{C}\mathbf{A}\mathbf{C}' = \mathrm{W}_m \sum_{\alpha=1}^{m} \mathbf{C}\underset{\sim}{\mathbf{Z}}_{\alpha}\underset{\sim}{\mathbf{Z}}'_{\alpha}\mathbf{C}' = \sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{Y}}_{\alpha}\underset{\sim}{\mathbf{Y}}'_{\alpha}$$

where $\underset{\sim}{\mathbf{Y}}_{\alpha} = \mathbf{C}\,\underset{\sim}{\mathbf{Z}}_{\alpha} \sim N_p\left(\underset{\sim}{\mathbf{0}}, \mathbf{C}\,\boldsymbol{\Sigma}\,\mathbf{C}'\right)$

[$\because E\left(\underset{\sim}{\mathbf{Y}}_{\alpha}\right) = CE\left(\underset{\sim}{\mathbf{Z}}_{\alpha}\right) = \underset{\sim}{\mathbf{0}}$

$V\left(\underset{\sim}{\mathbf{Y}}_{\alpha}\right) = V\left(\mathbf{C}\underset{\sim}{\mathbf{Z}}_{\alpha}\right) = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'$ and $\underset{\sim}{\mathbf{Y}}_{\alpha}$ is normal random vector]

$$\mathbf{C}\mathbf{A}\mathbf{C}' \sim \mathrm{W}_m\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)$$

Hence the proof.

P.D.F. Of Wishart Distribution :-

The p.d.f. of $\mathbf{A} \sim \mathrm{W}_n\left(\boldsymbol{\Sigma}\right)$ is given by

$$\mathrm{W}_n\left(\frac{\mathbf{A}}{\boldsymbol{\Sigma}}\right) = \frac{\left|\mathbf{A}\right|^{(n-p-1)/2} \, e^{tr\left(\mathbf{A}\boldsymbol{\Sigma}^{-1}\right)/2}}{2^{np/2}\,\pi^{p(p-1)/4}\left|\boldsymbol{\Sigma}\right|^{n/2}\prod_{i=1}^{p}\left|\left(\frac{1}{2}\left(n+1-i\right)\right)\right|}$$

$\mathbf{A}$ is positive definite and $\Gamma(.)$ is gamma function.

INFERENCES ABOUT MULTIVARIATE NORMAL MEAN VECTOR(S)

One Sample Problem:

Suppose $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ is a random sample from a multivariate normal population. Now, our statistical problem is whether the given sample has come from the multivariate normal population, whose mean vector is given by $\underset{\sim}{\mu} = \underset{\sim}{\mu}_0$. In other words, we have to test

$$H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0 \quad \text{vs} \quad H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$$

based on the given random sample $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$.

Two Sample Problem:

Suppose we have two different samples from two different multivariate normal populations $N_p(\underset{\sim}{\mu}^{(1)}, \Sigma)$ and $N_p(\underset{\sim}{\mu}^{(2)}, \Sigma)$ with common variance-covariance matrix Σ. Now, our statistical problem is whether the two normal populations have the same mean vector or not. In other words, our problem is equivalent to test the hypothesis

$$H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)} \quad \text{vs} \quad H_1 : \underset{\sim}{\mu}^{(1)} \neq \underset{\sim}{\mu}^{(2)}$$

based on the given two samples.

For developing the test statistics in the above two problems, we have to consider whether the common covariance matrix Σ is known or not. First, let us develop the test statistics for the above one-sample case as well as two-sample case assuming the population variance-covariance matrix Σ is known.

Developing Test Statistics When Σ is known :

One-Sample problem:

If $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ is a random sample of size *n* drawn from a multivariate normal population with known variance-covariance matrix Σ, then obtain the test statistic for testing

$$H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0 \quad \text{vs} \quad H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$$

and the critical region of size 'α' as well as the confidence region for $\underset{\sim}{\mu}$ of confidence 1- α.

Proof :-

In order to obtain the above result, let us prove the following theorem.

Theorem(1) :

If a p-component vector $\underset{\sim}{Y} \sim N_p(\underset{\sim}{0}, \Sigma)$, where Σ is non- singular (positive definite), then

$$\underset{\sim}{Y}' \Sigma^{-1} \underset{\sim}{Y} \sim \chi^2_p \qquad \rightarrow (1)$$

where χ^2_p is Chi-square distribution with *p* d.f .

Proof: We have given $\underset{\sim}{Y} \sim N_p(\underset{\sim}{0}, \Sigma)$.

Since, Σ is p.d.f $\exists$ a non-singular matrix **C** such that,

$$\mathbf{C \Sigma C' = I}$$

$$\Rightarrow \Sigma = \mathbf{C^{-1}\, I\,(C')^{-1}} = \left(\mathbf{C'\,C}\right)^{-1} \qquad \rightarrow (2)$$

Let us define the linear transformation,

$$\underset{\sim}{\mathbf{Z}} = \mathbf{C}\underset{\sim}{\mathbf{Y}} \qquad \rightarrow (3)$$

Then, $E(\underset{\sim}{Z}) = C\,E(\underset{\sim}{Y}) = \underset{\sim}{0}$

$V(\underset{\sim}{\mathbf{Z}}) = V(\mathbf{C}\underset{\sim}{\mathbf{Y}}) = C\,V(\underset{\sim}{\mathbf{Y}})\,C' = \mathbf{C\,\Sigma\,C'} = I$ (from (2))

Since the transformation is linear,

$\underset{\sim}{Z} \sim N_p(\underset{\sim}{0}, I)$ i.e ., $Z_1, Z_2, ..., Z_p$, the individual components of $\underset{\sim}{Z}$ are distributed as N(0,1).

Further, since the covariances are zeros, $Z_1, Z_2, ..., Z_p$ are independent which follows from the normality of the components.

$\therefore \quad \underset{\sim}{Z}'\underset{\sim}{Z} = Z_1^2 + Z_2^2 + ... + Z_p^2 \sim \chi_p^2$

$\Rightarrow \underset{\sim}{Y}'C'C\underset{\sim}{Y} \sim \chi_p^2$ (from (3))

$\Rightarrow \underset{\sim}{Y}'\Sigma^{-1}\underset{\sim}{Y} \sim \chi_p^2$ (from (2))

Hence the result (1) .

Proof of the Result:-

We have given the random sample $\mathbf{\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n}$ from $N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is known .

Now, we know that, the sample mean, $\bar{\underset{\sim}{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}/n)$.

Define the random vector, $\underset{\sim}{Y} = \sqrt{n}\,(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu})$ $\rightarrow (4)$

With $E(\underset{\sim}{Y}) = \sqrt{n}\,E(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}}) = \sqrt{n}\,(\underset{\sim}{\mu} - \underset{\sim}{\mu}) = \underset{\sim}{0}$.

$V(\underset{\sim}{\mathbf{Y}}) = V(\sqrt{n}\,(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}})) = n\,V(\bar{\underset{\sim}{\mathbf{X}}}) = n\,\Sigma/n = \Sigma$.

Thus, the mean vector of $\underset{\sim}{Y}$ is $\underset{\sim}{0}$ and covariance matrix is Σ.

Further, since the transformation in (4) is linear,

$$\underset{\sim}{Y} \sim N_p(\underset{\sim}{0}, \Sigma)$$

Now, from the above **theorem (1)**, it immediately follows

$$\underset{\sim}{Y}'\Sigma^{-1}\underset{\sim}{Y} \sim \chi_p^2$$

$$\Rightarrow \quad \sqrt{n}\,(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}})'\,\boldsymbol{\Sigma}^{-1}\,(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}})\,\sqrt{n} \sim \chi_p^2 \qquad \text{(from (4))}$$

$$\Rightarrow \quad n\,(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}})'\,\boldsymbol{\Sigma}^{-1}\,(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}}) \sim \chi_p^2$$

Thus, the test statistic for $H_0 : \underset{\sim}{\boldsymbol{\mu}} = \underset{\sim}{\boldsymbol{\mu}_0}$ is given by

$$n\,(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}}_0)'\,\boldsymbol{\Sigma}^{-1}\,(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}}_0) \qquad \rightarrow (5)$$

which follows χ^2 distribution with p d.f.

Let $\chi_p^2(\alpha)$ be the number such that $\Pr\left\{\chi_p^2 \geq \chi_p^2(\alpha)\right\} = \alpha$.

Thus, $\Pr\{n(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}}_0)'\,\boldsymbol{\Sigma}^{-1}\,(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}}_0) \geq \chi_p^2(\alpha)\} = \alpha$

and to test $H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ (given), we use

$$n(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}}_0)'\,\boldsymbol{\Sigma}^{-1}\,(\bar{\underset{\sim}{\mathbf{X}}} - \underset{\sim}{\boldsymbol{\mu}}_0) \geq \chi_p^2(\alpha) \qquad \rightarrow (6)$$

as critical region.

Similarly, we use the inequality,

$$(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}^*)' \, \Sigma^{-1} \, (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}^*) \leq \chi^2_p(\alpha) \qquad \rightarrow (7)$$

for obtaining the confidence region for $\underset{\sim}{\mu}$ (the set of all $\underset{\sim}{\mu}^*$ satisfying (7)) with confidence $1 - \alpha$.

Hence the result.

Two Sample Problem :-

Suppose we have a sample $\underset{\sim}{X}^{(1)}_1, \underset{\sim}{X}^{(1)}_2, ..., \underset{\sim}{X}^{(1)}_{n_1}$ from $N_p(\underset{\sim}{\mu}^{(1)}, \Sigma)$ and another sample $\underset{\sim}{X}^{(2)}_1, \underset{\sim}{X}^{(2)}_2, ..., \underset{\sim}{X}^{(2)}_{n_2}$ from $N_p(\underset{\sim}{\mu}^{(2)}, \Sigma)$, where Σ is known. Now, under the null hypothesis

$$H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)}$$

$$\frac{n_1 n_2}{n_1 + n_2} \left[\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} \right)' \Sigma^{-1} \left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} \right) \right] \sim \chi^2_p,$$

where, $\bar{\underset{\sim}{X}}^{(1)}$ = mean of the random sample $\underset{\sim}{X}^{(1)}_1, \underset{\sim}{X}^{(1)}_2, ..., \underset{\sim}{X}^{(1)}_{n_1}$

and $\bar{\underset{\sim}{X}}^{(2)}$ = mean of the random sample $\underset{\sim}{X}^{(2)}_1, \underset{\sim}{X}^{(2)}_2, ..., \underset{\sim}{X}^{(2)}_{n_2}$.

Solution:

From the given hypothesis, we have

$$\bar{\underset{\sim}{X}}^{(1)} \sim N_p\left(\underset{\sim}{\mu}^{(1)}, \Sigma/n_1\right) \;\&\; \bar{\underset{\sim}{X}}^{(2)} \sim N_p\left(\underset{\sim}{\mu}^{(2)}, \Sigma/n_2\right) \qquad \rightarrow (1)$$

Now define, $\underset{\sim}{Y} = \bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} \qquad \rightarrow (2)$

with mean vector, $E(\underset{\sim}{Y}) = \underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)} \qquad \rightarrow (3)$

and variance- covariance matrix,

$$\begin{aligned} V(\underset{\sim}{Y}) &= V(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}) \\ &= V(\bar{\underset{\sim}{X}}^{(1)}) + V(\bar{\underset{\sim}{X}}^{(2)}) - Cov(\bar{\underset{\sim}{X}}^{(1)}, \bar{\underset{\sim}{X}}^{(2)}) - Cov(\bar{\underset{\sim}{X}}^{(2)}, \bar{\underset{\sim}{X}}^{(1)}) \\ &= \frac{\Sigma}{n_1} + \frac{\Sigma}{n_2} = \Sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \qquad \rightarrow (4) \end{aligned}$$

(since the two samples are independent and as a consequence the covariance matrices $Cov(\bar{\underset{\sim}{X}}^{(1)}, \bar{\underset{\sim}{X}}^{(2)}) = \underset{\sim}{0} \;\&\; Cov(\bar{\underset{\sim}{X}}^{(2)}, \bar{\underset{\sim}{X}}^{(1)}) = \underset{\sim}{0}$

Since the transformation used in (2) is linear, we have

$$\underset{\sim}{Y} \sim N_p\left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}, \Sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

$$\Rightarrow \quad \underset{\sim}{Y} - (\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}) \sim N_p\left(\underset{\sim}{0}, \Sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Now, from the above theorem (1), it immediately follows

$$\left[\underset{\sim}{Y} - (\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)})\right]' \left[\Sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{-1} \left[\underset{\sim}{Y} - (\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)})\right] \sim \chi^2_p$$

But, under the null hypothesis, $H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)}$

$$\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right)' \left[\Sigma\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right]^{-1} \left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) \sim \chi^2_p \quad \text{(from(2))}$$

$$\Rightarrow \frac{n_1 n_2}{n_1+n_2}\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right)' \Sigma^{-1} \left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) \sim \chi^2_p$$

Hence the proof .

NOTE :-

(i) The critical region for testing hypothesis, $H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)}$ at $\boldsymbol{\alpha}$ level of significance is given by $\left(\frac{n_1 n_2}{n_1+n_2}\right)\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right)' \Sigma^{-1} \left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) \geq \chi^2_p(\alpha)$

where, $\chi^2_p(\alpha)$ is a number such that $P\left(\chi^2_p \leq \chi^2_p(\alpha)\right) = \alpha$.

(ii) Mahalonobis (1930) suggested , $(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)})' \Sigma^{-1} (\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)})$ as a measure of the distance between two populations .

(iii) The $(1-\alpha)\%$ confidence interval for $\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}$ is given by the set of $\underset{\sim}{\nu}^*$ all which are satisfying the inequality,

$$\left(\frac{n_1 n_2}{n_1+n_2}\right)\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} - \underset{\sim}{\nu}^*\right)' \Sigma^{-1} \left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} - \underset{\sim}{\nu}^*\right) \leq \chi^2_p(\alpha) .$$

When 'Σ'Is Unknown:-

Introduction:-

One of the most important groups of problems in Univariate theory relates to questions concerning the mean (μ) of a give distribution, when the variance (σ^2) of the distribution is unknown . On the basis of a sample, one may wish to decide whether the mean is equal to a specified number (μ_0) or one may wish to give an interval with in which 'μ' lies . The following is the mathematical formulation of one sample t-test.

Suppose $X_1, X_2, \ldots\ldots\ldots, X_n$ denote a random sample from a normal population $N(\mu, \sigma)$. Now the well – known test statistic for testing the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ is given by

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \quad \text{where,} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad , \; S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 .$$

This test statistic follows students' t-distribution with n-1 degrees of freedom . We reject H_0 is $|t|$ exceeds a specified percentage point of a t-distribution with n-1 degrees of freedom .

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$(\bar{X} - \mu) \sim N\left(0 , \sigma^2/n\right)$$

i.e., $\sqrt{n}\,(\bar{X} - \mu) \sim N(0 , \sigma)$

Regecting H_0 when $|t|$ is large is equivalently to rejecting H_0 is its square.

$$t^2 = \frac{(\bar{X} - \mu_0)^2}{s^2/n} = n(\bar{X} - \mu_0)(S^2)^{-1}(\bar{X} - \mu_0) \quad \rightarrow (1)$$

is large . Thus the test becomes .

Reject H_0 in favour of H_1 at α level of significance, if

$$n(\bar{X} - \mu_0)(S^2)^{-1}(\bar{X} - \mu_0) > t^2_{n-1}(\alpha/2) \quad \rightarrow (2)$$

where, $t_{n-1}(\alpha/2)$ denotes the upper $100(\alpha/2)^{th}$ percentile of the t-distribution with n-1d.f and $\bar{X}$ is the mean of the given sample $x_1, x_2, \ldots\ldots, x_n$ and S^2 is the sample variance of the given sample.

Now, the multivariate analogue of t^2 given in (1) is

$$T^2 = n(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}_0)' S^{-1}(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}_0) \quad \rightarrow (3)$$

Where, $\underset{\sim}{\bar{X}}$ & S are respectively the sample mean and sample covariance matrix obtained . Using the multivariate normal sample $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, \ldots, \underset{\sim}{\mathbf{X}}_n$ and are given by

$$\underset{\sim}{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{X}_i$$

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\underset{\sim}{\mathbf{X}}_i - \underset{\sim}{\bar{\mathbf{X}}})(\underset{\sim}{\mathbf{X}}_i - \underset{\sim}{\bar{\mathbf{X}}})' \quad \rightarrow (4)$$

and $\underset{\sim}{\mu}_0 = \begin{pmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{pmatrix}_{p\times 1}$ is the specified value of population mean vector $\underset{\sim}{\mu}$.

The statistic T^2 is called Hotelling's T^2 in honour of Harnold Hotelling, a pioneer in multivariate analysis, who first obtained its sampling distribution.

It may be proved that $T^2 \sim \frac{(n-1)p}{(n-p)} F_{p,n-p} \quad \rightarrow (5)$

where, $F_{p,n-p}$ denotes F-distribution with p, n-p degrees of freedom.

The above explanation about T^2 statistic can be summarized as follows.

Let $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, \ldots, \underset{\sim}{\mathbf{X}}_n$ be a random sample from $N_p(\underset{\sim}{\mu}, \Sigma)$ population (Σ is unknown), then Hotellings T^2– stastistics for testing.

$H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ against $H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$, is given by (3) and H_0 may be selected at α-level of significance infevour of H_1 is

$$T^2 = n(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}_0)' S^{-1}(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}_0)$$

$$> \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha) \quad \rightarrow (6)$$

where, $F_{p,n-p}(\alpha)$ is the upper $100\alpha^{th}$ percentile of the $F_{p,n-p}$ distribution and $\underset{\sim}{\bar{X}}$ & S are respectively . The mean vector and covariance matrix of the given sample $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$.

Definition of Hotelling's T^2**-distribution(statistic)**:

Suppose $\underset{\sim}{Y}$ is a *p*-variate random vector distributed according to $N_p(\underset{\sim}{0}, \Sigma)$.

Suppose $\mathbf{B} = \sum_{i=1}^{n} \underset{\sim}{Z}_i \underset{\sim}{Z}_i'$ (where each $\underset{\sim}{Z}_i \sim N_p(\underset{\sim}{0}, \Sigma)$ and are independent) is a Wishart random matrix and is distributed as Wishart distribution with n degrees of freedom i.e. $\mathbf{B} \sim \mathbf{W_n(\Sigma)}$.

Now, if $\underset{\sim}{Y}$ and $\mathbf{B}$ are independent then the quantity

$$T^2 = \underset{\sim}{Y}' \left(\frac{\mathbf{B}}{n} \right)^{-1} \underset{\sim}{Y}$$

is called as Hotelling's T^2 statistic and it distribution is called as Hotelling's T^2-distribution with n d.f. and is denoted as

$$T^2 \sim T_n^2$$

NOTE :-

(1) the statistic $\frac{T^2}{n}$ is the ratio of two independent χ^2 - variates with p d.f. and n-p+1 d.f. respectively i.e., $\frac{T^2}{n} = \frac{\chi_p^2}{\chi_{n-p+1}^2}$.

(2) $\frac{T^2}{n}\left(\frac{n - p + 1}{p} \right) \sim F_{p,n-p+1}$, where $F_{p,n-p+1}$ is the F-distribution with p and n-p+1 degrees of freedom.

THE LIKELIHOOD RATION PRINCIPLE

Deriving T^2-statistic as the Likelihood Ratio Test of $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu_0}$:

There is a general principle for constructing test procedures called the Likelihood Ratio (LR) principle method and the T^2-statistic can be derived as the LR test of $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu_0}$ as explained below.

Suppose $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ (n > p) is given random sample from $N_p(\underset{\sim}{\mu}, \Sigma)$, the likelihood function is

$$L(\underset{\sim}{\mu},\Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{\alpha=1}^{n}(\underset{\sim}{x}_\alpha-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}_\alpha-\underset{\sim}{\mu})} \rightarrow (1)$$

Under the hypothesis, $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$, the likelihood becomes,

$$L(\underset{\sim}{\mu}_0,\Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{\alpha=1}^{n}(\underset{\sim}{x}_\alpha-\underset{\sim}{\mu}_0)'\Sigma^{-1}(\underset{\sim}{x}_\alpha-\underset{\sim}{\mu}_0)} \rightarrow (2)$$

The likelihood ratio criterion is

$$\lambda = \frac{\max_{\Sigma} L(\underset{\sim}{\mu}_0,\Sigma)}{\max_{\underset{\sim}{\mu},\Sigma} L(\underset{\sim}{\mu},\Sigma)} \rightarrow (3)$$

i.e., the numerator is the maximum of the likelihood function for $\underset{\sim}{\mu}, \Sigma$ is the parameter space restricted by the null hypothesis ($\underset{\sim}{\mu} = \underset{\sim}{\mu}_0$) and Σ is positive definite and the denominator is the maximum over the entire parameter space (Σ is positive definite).

When the parameters are unrestricted the MLE's of $\underset{\sim}{\mu}$ *and* Σ from (1) are given by

$$\hat{\underset{\sim}{\mu}}_\Omega = \bar{\underset{\sim}{x}}$$

$$\hat{\Sigma}_\Omega = \frac{1}{n}\sum_\alpha (\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})' \rightarrow (4)$$

When $\underset{\sim}{\mu} = \underset{\sim}{\mu}_0$, the likelihood function given by (2), minimizes at

$$\hat{\Sigma}_0 = \frac{1}{n}\sum_\alpha (\underset{\sim}{x}_\alpha - \underset{\sim}{\mu}_0)(\underset{\sim}{x}_\alpha - \underset{\sim}{\mu}_0)' \rightarrow (5)$$

Substituting (4) in (1), we get (after simplification),

$$\max_{\underset{\sim}{\mu},\Sigma} L(\underset{\sim}{\mu},\Sigma) = \frac{1}{(2\pi)^{np/2}\left|\hat{\Sigma}_\Omega\right|^{n/2}} e^{-np/2} \rightarrow (6)$$

Similarly, substituting (5) in (2),we get

$$\max_{\Sigma} L(\underset{\sim}{\mu}_0,\Sigma) = \frac{1}{(2\pi)^{np/2}\left|\hat{\Sigma}_0\right|^{n/2}} e^{-np/2} \rightarrow (7)$$

Substituting (6) & (7) in (3),we get,

$$\lambda = \left(\frac{\left|\hat{\Sigma}_\Omega\right|}{\left|\hat{\Sigma}_0\right|}\right)^{n/2}$$

$$\Rightarrow \quad \lambda^{2/n} = \frac{\left|\hat{\Sigma}_{\Omega}\right|}{\left|\hat{\Sigma}_{0}\right|} = \frac{\left|\sum_{\alpha=1}^{n}(\underset{\sim}{x}_{\alpha} - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_{\alpha} - \underset{\sim}{\bar{x}})'\right|}{\left|\sum_{\alpha=1}^{n}(\underset{\sim}{x}_{\alpha} - \underset{\sim}{\mu}_{0})(\underset{\sim}{x}_{\alpha} - \underset{\sim}{\mu}_{0})'\right|} \quad \rightarrow (7a)$$

$$\Rightarrow \quad \lambda^{2/n} = \frac{\left|\mathbf{A}\right|}{\left|\mathbf{A} + n\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})'\right|}$$

Where $\mathbf{A} = \sum_{\alpha=1}^{n}(\underset{\sim}{x}_{\alpha} - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_{\alpha} - \underset{\sim}{\bar{x}})' \quad \rightarrow (8)$

Consider the matrix, $\underset{(p+1)\times(p+1)}{\mathbf{B}} = \begin{bmatrix} \mathbf{A} & \sqrt{n}(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0}) \\ \sqrt{n}(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})' & -1 \end{bmatrix}$

$$= \begin{bmatrix} \mathbf{B}_{11} & \vdots & \mathbf{B}_{12} \\ \cdots & \vdots & \cdots \\ \mathbf{B}_{21} & \vdots & \mathbf{B}_{22} \end{bmatrix}.$$

We have, $\left|\mathbf{B}\right| = \left|\mathbf{B}_{11}\right|\left|\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\right|$

$$= \left|\mathbf{B}_{22}\right|\left|\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21}\right|$$

$$\therefore \quad \left|\mathbf{B}\right| = \left|-1\right| \left|\mathbf{A} - \sqrt{n}\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})\,(-1)^{-1}\sqrt{n}\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})'\right|$$

$$= \left|\mathbf{A} + n\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})'\right|.$$

$$\therefore \quad \lambda^{2/n} = \frac{\left|\mathbf{A}\right|}{-\begin{vmatrix} \mathbf{A} & \sqrt{n}\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0}) \\ \sqrt{n}\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})' & -1 \end{vmatrix}}$$

$$= \frac{\left|\mathbf{A}\right|}{-\left|\mathbf{A}\right| \left|-1 - \sqrt{n}\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})'\,\mathbf{A}^{-1}\sqrt{n}\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})\right|}$$

$$= \frac{\left|\mathbf{A}\right|}{\left|\mathbf{A}\right| \left|1 + n\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})'\,\mathbf{A}^{-1}\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})\right|}$$

$$= \frac{1}{1 + n(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})'\,\mathbf{A}^{-1}\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_{0})}.$$

Where $\mathbf{A}$ is as given in (8) . But we have,

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A} = \frac{1}{n-1}\sum_{\alpha}(\underset{\sim}{x}_{\alpha} - \underset{\sim}{\bar{x}})\,(\underset{\sim}{x}_{\alpha} - \underset{\sim}{\bar{x}})'$$

$$\Rightarrow \quad \mathbf{A} = (n-1)\,\mathbf{S}$$

$$\therefore\ \lambda^{2/n} = \frac{1}{1+n(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu_0})' [(n-1)S]^{-1} (\underset{\sim}{\bar{x}} - \underset{\sim}{\mu_0})}$$

$$= \frac{1}{1+\frac{n}{(n-1)}(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu_0})' S^{-1} (\underset{\sim}{\bar{x}} - \underset{\sim}{\mu_0})} = \frac{1}{1+T^2/(n-1)} \qquad \rightarrow (9)$$

where, $T^2 = n(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu_0})' S^{-1} (\underset{\sim}{\bar{x}} - \underset{\sim}{\mu_0})$ is Hotelling's T^2-statistic.

Now, from (7a) & (9), we can see

$$1 + \frac{T^2}{(n-1)} = \frac{\left|\sum_{\alpha=1}^{n} (\underset{\sim}{x_\alpha} - \underset{\sim}{\mu_0})(\underset{\sim}{x_\alpha} - \underset{\sim}{\mu_0})'\right|}{\left|\sum_{\alpha=1}^{n} (\underset{\sim}{x_\alpha} - \underset{\sim}{\bar{x}})(\underset{\sim}{x_\alpha} - \underset{\sim}{\bar{x}})'\right|}$$

$$\Rightarrow \quad T^2 = (n-1)\left[\frac{\left|\sum_{\alpha=1}^{n} (\underset{\sim}{x_\alpha} - \underset{\sim}{\mu_0})(\underset{\sim}{x_\alpha} - \underset{\sim}{\mu_0})'\right|}{\left|\sum_{\alpha=1}^{n} (\underset{\sim}{x_\alpha} - \underset{\sim}{\bar{x}})(\underset{\sim}{x_\alpha} - \underset{\sim}{\bar{x}})'\right|} - 1\right]$$

$$= (n-1)\left[\frac{\left|\hat{\Sigma}_0\right|}{\left|\hat{\Sigma}_\Omega\right|} - 1\right] \qquad \rightarrow (10)$$

In this formula, we need not find the inverse of a matrix, where as in the original formula we have to evaluate S^{-1} .

Theorem : T^2 - statistic is invariant (unchanged) under changes in the units of measurements for $\underset{\sim}{X}$ of the form,

$$\underset{\sim}{Y} = C\underset{\sim}{X} + \underset{\sim}{d}, \text{ where C is non-singular} \qquad \rightarrow (1)$$

Proof :- We have, $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$,

$$\text{i.e., } E(\underset{\sim}{X}) = \underset{\sim}{\mu} \ \Rightarrow \ E(\underset{\sim}{Y}) = C\underset{\sim}{\mu} + \underset{\sim}{d} \quad (\because \text{from (1)}) \qquad \rightarrow (2)$$

Now, we have the T^2 - statistic for testing, $H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu_0}$ vs $H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu_0}$ based on the given sample $\underset{\sim}{X_1}, \underset{\sim}{X_2}, ..., \underset{\sim}{X_n}$ is

$$T_x^2 = n\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu_0})' S_x^{-1} (\underset{\sim}{\bar{x}} - \underset{\sim}{\mu_0}) \qquad \rightarrow (3)$$

where $$S_x = \frac{1}{n-1}\sum_{i=1}^{n} (\underset{\sim}{x_i} - \underset{\sim}{\bar{x}})(\underset{\sim}{x_i} - \underset{\sim}{\bar{x}})' \qquad \rightarrow (3a)$$

From (1) we can see $\underset{\sim}{Y} \sim N_p(C\underset{\sim}{\mu} + \underset{\sim}{d}, C\Sigma C')$.

Now, the T^2 - statistic for testing,

$H_0 : \underset{\sim}{\mu_Y} = \underset{\sim}{\mu_{Y_0}}$ vs $H_1 : \underset{\sim}{\mu_Y} \neq \underset{\sim}{\mu_{Y_0}}$

where, $\underset{\sim}{\mu_Y} = C\underset{\sim}{\mu} + \underset{\sim}{d}$ & $\underset{\sim}{\mu_{Y_0}} = C\underset{\sim}{\mu_0} + \underset{\sim}{d}$ $\qquad \rightarrow (4)$

based on the sample $\underset{\sim}{Y_1}, \underset{\sim}{Y_2}, ..., \underset{\sim}{Y_n}$ is given by

$$T_y^2 = n\,(\underset{\sim}{\bar{Y}} - \underset{\sim}{\mu}_{Y_0})'\, S_Y^{-1} (\underset{\sim}{Y} - \underset{\sim}{\mu}_{Y_0}) \qquad \rightarrow (5)$$

where, $\underset{\sim}{\bar{Y}} = C\,\underset{\sim}{\bar{x}} + \underset{\sim}{d}$ (from (1))

$$S_y = \frac{1}{n-1} \sum_{i=1}^{n} (\underset{\sim}{y}_i - \underset{\sim}{\bar{y}})(\underset{\sim}{y}_i - \underset{\sim}{\bar{y}})' \qquad \rightarrow (6)$$

and $\underset{\sim}{\mu}_{Y_0}$ is given by (4).

In order to show that the Hotelling's T^2 is invariant under the changes in the units of measurements, we have to show, $T_y^2 = T_x^2$.

For that, consider T_y^2 given from (5),

$$\begin{aligned} T_y^2 &= n\,(\underset{\sim}{\bar{y}} - \underset{\sim}{\mu}_{y_0})'\, S_y^{-1}\, (\underset{\sim}{\bar{y}} - \underset{\sim}{\mu}_{y_0}) \\ &= n\,(C\underset{\sim}{\bar{x}} - C\underset{\sim}{\mu}_0)'\, S_y^{-1}\, (C\underset{\sim}{\bar{x}} - C\underset{\sim}{\mu}_0) \qquad \text{(using (4))} \\ &= n\,(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_0)'\, C'\, S_y^{-1}\, C(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_0) \qquad \rightarrow (7) \end{aligned}$$

But, $S_y = \frac{1}{n-1} \sum_{i=1}^{n} (\underset{\sim}{y}_i - \underset{\sim}{\bar{y}})(\underset{\sim}{y}_i - \underset{\sim}{\bar{y}})'$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (C\underset{\sim}{x}_i - C\underset{\sim}{\bar{x}})(C\underset{\sim}{x}_i - C\underset{\sim}{\bar{x}})' \qquad \text{(using (1))}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} C(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}})' C'$$

$$= CS_xC' \qquad \text{(from (3a))}$$

$$\Rightarrow \; S_y^{-1} = (C')^{-1} S_x^{-1} C^{-1}$$

$$\Rightarrow \; C' S_y^{-1} C = S_x^{-1}$$

Using this in (7), we get

$$T_y^2 = n(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_0)' S_x^{-1} (\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}_0) = T_x^2 \qquad \text{(from (3))}$$

Thus, T^2 is invariant under the changes in the units of measurements.

NOTE : The above theorem may be stated as " The Hotellings T^2 is invariant under linear transformation (or under changes in the location and scale) of the sample .

Uses Of The T^2-statistics :-

(i)**For testing the significance of one sample mean vector $\underset{\sim}{\bar{x}}$:**

Suppose $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ is a random sample from a *p*-variate normal population $N_p(\underset{\sim}{\mu}, \Sigma)$, where both $\underset{\sim}{\mu}$ and Σ are assumed as unknown. Now, our statistical problem is whether the given sample has come from the multivariate normal population, whose mean vector is $\underset{\sim}{\mu}_0$. In other words, we want to test the hypothesis

$$H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0 \;\text{ vs }\; H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0 \qquad \rightarrow (1),$$

where $\underset{\sim}{\mu}_0$ is the given mean vector.

For testing the above hypothesis, derive the test statistic.

Solution:

We have given a random sample of size n viz., $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ from $N_p(\underset{\sim}{\mu}, \Sigma)$, where both $\underset{\sim}{\mu}$ and Σ are unknown.

Now, we know that the mean vector

$$\bar{\underset{\sim}{X}} \sim N_p(\underset{\sim}{\mu}, \Sigma/n) \quad \text{(since } \bar{\underset{\sim}{X}} \text{ is a linear function of the sample)}$$

Define the random vector, $\underset{\sim}{Y} = \sqrt{n}(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}) \rightarrow (2)$

Whose population mean vector and population variance-covariance matrix are respectively given by

$$E(\underset{\sim}{Y}) = \sqrt{n}E(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}) = \sqrt{n}(\underset{\sim}{\mu} - \underset{\sim}{\mu}) = \underset{\sim}{0}.$$

$$V(\underset{\sim}{Y}) = V(\sqrt{n}(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu})) = n\, V(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}) = n\, V(\bar{\underset{\sim}{X}}) = n\, \Sigma/n = \Sigma.$$

Thus, the mean vector of $\underset{\sim}{Y}$ is $\underset{\sim}{0}$ and covariance matrix is Σ.

Further, since the transformation in (2) is linear, we have

$$\underset{\sim}{Y} \sim N_p(\underset{\sim}{0}, \Sigma) \rightarrow (3)$$

We have the sample variance-covariance matrix

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})' \rightarrow (4)$$

Now, we know that $(n-1)S$ follows Wishart distribution with n-1 degrees freedom and parameter Σ that is

$$(n-1)S \sim W_{n-1}(\Sigma) \rightarrow (5)$$

Further, we know that the sample mean vector $\bar{\underset{\sim}{X}}$ and the sample variance-covariance matrix S are independently distributed.

From (2), it immediately follows that the random vector $\underset{\sim}{Y}$ and the random matrix $(n-1)S$ distribute independently.

Now, by the definition of Hotelling's T^2 distribution, the statistic

$$T^2 = \underset{\sim}{Y}'\left(\frac{(n-1)S}{(n-1)}\right)^{-1}\underset{\sim}{Y} \rightarrow (6)$$

follows Hotelling's T^2 distribution with n-1d.f. i.e.

$$T^2 \sim T^2_{n-1}$$

Substituting (2) in (6), we can see that

$$T^2 = n(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu})'S^{-1}(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}) \sim T^2_{n-1} \rightarrow (7)$$

Now, under $H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ (7) becomes

$$T^2 = n(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}_0)'S^{-1}(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}_0) \sim T^2_{n-1} \rightarrow (8)$$

where, $\bar{\underset{\sim}{X}} = \frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{X}_i$ and $S = \frac{1}{n-1}\sum_{i=1}^{n}(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})'$

Thus, the formula (8) gives us the Hotelling's T^2 statistic which can be used to test (1) and follows T^2_{n-1}

At the given α level of significance, H_0 may be rejected infavour of H_1 if

$$\frac{T^2}{n-1}\left(\frac{n-p}{p}\right) > F_{p,n-p}(\alpha) \quad \text{(or)} \quad T^2 > T_0^2 \ . \qquad \rightarrow (5)$$

Where, $T_0^2 = \frac{(n-1)\,p}{n-p}\ F_{p,n-p}\ (\alpha)$ and $F_{p,n-p}\ (\alpha)$ is the upper $100\ \alpha^{th}$ percentile of the F-distribution and can be obtained from the F-tables.

Nature Of T^2- statistic (Distribution) :-

We can write,

$$T^2 = \sqrt{n}\ (\underset{\sim}{\bar{X}} - \mu_0)' \left(\frac{\sum_{i=1}^{n} (\underset{\sim}{X}_i - \underset{\sim}{\bar{X}})\ (\underset{\sim}{X}_i - \underset{\sim}{\bar{X}}_i)'}{n-1} \right)^{-1} \sqrt{n}\ (\underset{\sim}{\bar{X}} - \mu_0)$$

which is of the form,

$$\begin{pmatrix} \text{multivariate} \\ \text{normal r.v} \end{pmatrix}' \left(\frac{\text{Wishart random matrix}}{\text{d.f}} \right)^{-1} \begin{pmatrix} \text{multivariate} \\ \text{normal r.v} \end{pmatrix}.$$

Since the multivariate normal random vector and the Wishart random matrix, given in T^2 are independently distributed ($\because \underset{\sim}{\bar{X}}$ & S are independently distributed). Their joint distribution is the product of the marginal normal and Wishart distributions and therefore T^2-distribution can be obtained from this.

(ii) **Computation Of Confidence Region For Mean Vector** :-

The $100(1-\alpha)$ percent confidence region for the population mean $\underset{\sim}{\mu}$ of $N_p(\underset{\sim}{\mu}, \Sigma)$ based on the given random sample $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ is given by the set of all $\underset{\sim}{\mu}$ satisfying the inequality,

$$n\ (\underset{\sim}{\bar{x}} - \underset{\sim}{\mu})'\ S^{-1}\ (\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}) \le T_0^2(\alpha) \ \text{,where } T_0^2(\alpha) = \frac{(n-1)\,p}{n-p}\ F_{p,n-p}\ (\alpha).$$

(iii) **A Two Sample Problem (An application of Hotelling T^2-statistic)** :-

Another situation in which the T^2-statistic is used is that in which the null hypothesis is that the mean of one normal population is equal to the mean of the other, where the covariance matrices are assumed equal but unknown.

Suppose $\underset{\sim}{X}_1^{(1)}, \underset{\sim}{X}_2^{(1)}, ..., \underset{\sim}{X}_n^{(1)}$ is a sample from $N_p(\underset{\sim}{\mu}^{(1)}, \Sigma)$ and $\underset{\sim}{X}_1^{(2)}, \underset{\sim}{X}_2^{(2)}, ..., \underset{\sim}{X}_n^{(2)}$ is a another sample (independent of the first sample) from $N_p(\underset{\sim}{\mu}^{(2)}, \Sigma)$.

Now, we wish to test the null hypothesis,

$H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)}$ or $\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)} = 0$,against $H_1 : \underset{\sim}{\mu}^{(1)} \neq \underset{\sim}{\mu}^{(2)}$ $\rightarrow (1)$

The sample means from the hypothesis ,

$$\bar{\underset{\sim}{X}}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} \underset{\sim}{X}_i^{(1)} \sim N_p\left(\underset{\sim}{\mu}^{(1)}, \Sigma / n_1\right)$$

and $$\bar{\underset{\sim}{X}}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2} \underset{\sim}{X}_i^{(2)} \sim N_p\left(\underset{\sim}{\mu}^{(2)}, \Sigma / n_2\right)$$

Now define, $\underset{\sim}{Y} = \bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}$ $\rightarrow (2)$

with mean, $E(\underset{\sim}{Y}) = E(\bar{\underset{\sim}{X}}^{(1)}) - E(\bar{\underset{\sim}{X}}^{(2)}) = \underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}$ $\rightarrow (3)$

and the variance- covariance matrix,

$V(\underset{\sim}{Y}) = V(\bar{\underset{\sim}{X}}^{(1)}) + V(\bar{\underset{\sim}{X}}^{(2)})$ ($\because$ The two samples are independent)

$$= \frac{1}{n_1}\Sigma + \frac{1}{n_2}\Sigma$$

$$= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma \qquad \rightarrow (4)$$

Since the transformation in (2) is linear, from (3) and (4) it follows

$$\underset{\sim}{Y} \sim N_p\left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}, \left(\frac{1}{n_1}+\frac{1}{n_2}\right)\Sigma\right)$$

i.e., $$\underset{\sim}{Y} - (\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}) \sim N_p\left(\underset{\sim}{0}, \left(\frac{1}{n_1}+\frac{1}{n_2}\right)\Sigma\right)$$

i.e., $$\frac{\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N_p(\underset{\sim}{0}, \Sigma) \quad \text{(using (2))} \quad \rightarrow (5)$$

The sample covariance matrix from sample 1, which is denoted by S_1 and is given by

$$S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} (\underset{\sim}{X}_i^{(1)} - \bar{\underset{\sim}{X}}^{(1)})(\underset{\sim}{X}_i^{(1)} - \bar{\underset{\sim}{X}}^{(1)})' .$$

Similarly, the sample covariance matrix from sample 2, denoted by S_2 and is given by

$$S_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2} (\underset{\sim}{X}_i^{(2)} - \bar{\underset{\sim}{X}}^{(2)})(\underset{\sim}{X}_i^{(2)} - \bar{\underset{\sim}{X}}^{(2)})' .$$

Let us denote, $S = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1 + n_2 - 2}$ $\rightarrow (6)$

We know that $(n_1\text{-}1)\mathbf{S}_1$ and $(n_2\text{-}1)\mathbf{S}_2$ are Wishart random matrices and are distributed as $w_{n_1\text{-}1}(\mathbf{\Sigma})$ & $w_{n_2\text{-}1}(\mathbf{\Sigma})$ respectively, where $w_{n_1\text{-}1}(\mathbf{\Sigma})$ is Wishart distribution with $(n_1 - 1)$ d.f and $w_{n_2\text{-}1}(\mathbf{\Sigma})$ is Wishart distribution with $(n_2 - 1)$ d.f. both have the parametric matrix Σ.

By assumption, the samples are independent, so $(n_1\text{-}1)\, S_1$ and $(n_2\text{-}1)\, S_2$ are also independent . Therefore from (6), $(n_1+n_2 - 2)S$ is distributed as Wishart distribution with $n_1+ n_2 - 2$ d.f and with the parametric matrix Σ, i.e. $(n_1+n_2 - 2)S \sim w_{n_1+ n_2\text{-}2}(\Sigma) \rightarrow (7)$

Since, the sample variance-covariance matrix is independently distributed with the sample mean vector, S_1 is independently distributed with $\underset{\sim}{\bar{X}}^{(1)}$ and since the two samples are independent, S_1 is independently distributed with $\underset{\sim}{\bar{X}}^{(2)}$ and therefore S_1 is independently distributed with $\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}$.

Similarly, S_2 is independently distributed with $\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}$.

Therefore, $S = \dfrac{(n_1 - 1)\, S_1 + (n_2 - 1)\, S_2}{n_1 + n_2 - 2}$ is independently distributed with $\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}$.

Thus, from the above explanation and from (5) & (6) and by the definition of T^2-distribution, we have

$$T^2 = \left[\frac{\left(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right]' S^{-1} \left[\frac{\left(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right]$$

$$= \left(\frac{n_1 n_2}{n_1+n_2}\right)\left[\left(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)\right]' \mathbf{S}^{-1}\left[\left(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)\right] \rightarrow (8)$$

is distributed as T^2-distribution with $n_1+ n_2 - 2$ d.f .

Now, by virtue of the relation between T^2 and F – distribution, we have

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

i.e., $$\frac{T^2}{n_1+ n_2 - 2} \sim \frac{p}{n_1+ n_2 - p\text{-}1} F_{p,\, n_1+ n_2 - p\text{-}1}$$

under $H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)}$ i.e., $\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)} = \underset{\sim}{0}$, (8) becomes

$$T^2 = \left(\frac{n_1 n_2}{n_1+n_2}\right)\left(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}\right) \mathbf{S}^{-1} \left(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}\right) \rightarrow (9)$$

if $T^2 > \frac{p(n_1+n_2-2)}{n_1+n_2-p-1} F_{p,\, n_1+n_2-p-1}(\alpha)$, where $F_{p,\, n_1+n_2-p-1}(\alpha)$ is table F-value at α level of significance with $(p,\ n_1+n_2-p-1)$ d.f., then $H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)}$ may be rejected .

Confidence Interval :-

An $100(1-\alpha)\%$ confidence region for estimation of $\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}$ is given by the following set ($\because$ from (8)),

$$\left\{ \underset{\sim}{m} \Big/ \left(\frac{n_1 n_2}{n_1+n_2} \right) \left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} - \underset{\sim}{m} \right) S^{-1} \left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} - \underset{\sim}{m} \right) < T_0^2 \right\}$$

where, $T_0^2 = \frac{p(n_1+n_2-2)}{n_1+n_2-p-1} F_{p,\, n_1+n_2-p-1}(\alpha)$.

The Two Sample Situation,When $\Sigma_1 \neq \Sigma_2$:-

In the above problem, we have assumed that the covariance matrices of both the populations are assumed as equal i.e., $\Sigma_1 = \Sigma_2 = \Sigma$. Now, let us suppose that $\Sigma_1 \neq \Sigma_2$ i.e., the population covariance matrices are not equal .

In this case, no tests are available for making inferences about $\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}$,when the sizes of the samples are small . However, if n_1 & n_2 are large i.e., in case of large samples,we have the following result.

Result :-

Let the sample sizes be such that n_1-p and n_2-p are large . An approximation $100(1-\alpha)\%$ confidence region for $\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}$ is given by all $\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}$ satisfying ,

$$\left[\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} \right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)} \right) \right]' \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right]^{-1} \left[\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} \right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)} \right) \right] \leq \chi_p^2(\alpha)$$

where, $\chi_p^2(\alpha)$ is χ^2- table values with p.d.f at $100\alpha\%$ level of significance.

Proof:- $E\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) = \underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}$

& $V\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) = V\left(\bar{\underset{\sim}{X}}^{(1)}\right) + V\left(\bar{\underset{\sim}{X}}^{(2)}\right) = \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2$.

By the central limit theorem,

$$\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} \sim N_p\left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}, \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right) .$$

If Σ_1 & Σ_2 are known,

$$\left[\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)\right]' \left[\frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right]^{-1}$$

$$\left[\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)\right] \sim \chi^2_p(\alpha),$$

approximately, when n_1 & n_2 are large, with high probability $S_1 \to \Sigma_1$ and $S_2 \to \Sigma_2$. Consequently, the approximation holds with S_1 & S_2, in place of Σ_1 and Σ_2 respectively .

Hence the theorem .

NULL DISTRIBUTION OF HOTELLING T^2

Statement:- Let $\underset{\sim}{\mathbf{Y}} \sim N_p(\underset{\sim}{\mathbf{v}}, \mathbf{\Sigma})$ and let $\mathbf{A}$ be a Wishart random matrix independently distributed as $\sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{Z}}_\alpha \underset{\sim}{\mathbf{Z}}'_\alpha$, where $\underset{\sim}{\mathbf{Z}}_\alpha$'s are i.i.d $\sim N_p(\underset{\sim}{\mathbf{0}}, \mathbf{\Sigma})$. Also let

$$T^2 = m\underset{\sim}{\mathbf{Y}}'\mathbf{A}^{-1}\underset{\sim}{\mathbf{Y}} \qquad \to (1)$$

then , $\frac{T^2}{m}\left(\frac{m - p + 1}{p}\right)$ is distributed as a non-central F with p and $m-p+1$ d.f. and non-centrality parameter $\underset{\sim}{\mathbf{v}}'\mathbf{\Sigma}^{-1}\underset{\sim}{\mathbf{v}}$. Further, if $\underset{\sim}{\mathbf{v}} = \underset{\sim}{\mathbf{0}}$, then

$$\frac{T^2}{m}\left(\frac{m - p + 1}{p}\right) \sim F_{p, m-p+1} \qquad \to (2)$$

and the distribution of T^2 is called T^2-distribution.

Proof :- Since $\mathbf{\Sigma}$ is positive definite, there exists a non-singular $\mathbf{C}$ such that

$$\mathbf{C\Sigma C}' = \mathbf{I_p} \text{ so that , } \mathbf{\Sigma} = (\mathbf{C}'\mathbf{C})^{-1} \qquad \to (3)$$

Define , $\underset{\sim}{\mathbf{Y}}^* = \mathbf{C}\underset{\sim}{\mathbf{Y}}$ and $\mathbf{A}^* = \mathbf{CAC}'$ $\qquad \to (4)$

We can see that , $\mathbf{E}(\underset{\sim}{\mathbf{Y}}^*) = \mathbf{C}\underset{\sim}{\mathbf{v}} = \underset{\sim}{\mathbf{v}}^*$ (say)

$$V(\underset{\sim}{\mathbf{Y}}^*) = V(\mathbf{C}\underset{\sim}{\mathbf{Y}}) = \mathbf{C\Sigma C}' = \mathbf{I} \qquad \text{(using (3))} \qquad \to (5)$$

Thus , $\underset{\sim}{\mathbf{Y}}^* \sim N_p(\underset{\sim}{\mathbf{v}}^*, \mathbf{I})$.

Since $\mathbf{A}$ is distributed as $\sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha}$, $\mathbf{A}^* = \mathbf{CAC}'$ is distributed as

$$\mathbf{C}\sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha}\mathbf{C}' = \sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{Z}}^*_{\alpha} \underset{\sim}{\mathbf{Z}}^{*\prime}_{\alpha} \qquad \rightarrow (6)$$

where , $\underset{\sim}{\mathbf{Z}}^*_{\alpha} = \mathbf{C}\underset{\sim}{\mathbf{Z}}_{\alpha} \sim N_p(\mathbf{C}\underset{\sim}{\mathbf{0}}, \mathbf{C\Sigma C}') = N_p(\underset{\sim}{\mathbf{0}}, \mathbf{I})$.

Eq (1) can be written as

$$\begin{aligned} T^2 &= m\underset{\sim}{\mathbf{Y}}'\mathbf{A}^{-1}\underset{\sim}{\mathbf{Y}} \\ &= m\underset{\sim}{\mathbf{Y}}'\mathbf{C}'(\mathbf{C}')^{-1}\mathbf{A}^{-1}(\mathbf{C})^{-1}\mathbf{C}\underset{\sim}{\mathbf{Y}} \\ &= m(\mathbf{C}\underset{\sim}{\mathbf{Y}})'(\mathbf{CAC}')^{-1}(\mathbf{C}\underset{\sim}{\mathbf{Y}}) \\ &= m\underset{\sim}{\mathbf{Y}}^{*\prime}\mathbf{A}^{*\prime}\underset{\sim}{\mathbf{Y}}^* \qquad \rightarrow (7) \end{aligned}$$

where, $\underset{\sim}{\mathbf{Y}}^* \sim N_p(\underset{\sim}{\mathbf{0}}, \mathbf{I})$ and $\mathbf{A}^*$ is independently distributed as $\sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{Z}}^*_{\alpha} \underset{\sim}{\mathbf{Z}}^{*\prime}_{\alpha}$ in which $\underset{\sim}{\mathbf{Z}}^*_{\alpha}$'s are i.i.d $\sim N_p(\underset{\sim}{\mathbf{0}}, \mathbf{I})$.

Also, since $\underset{\sim}{\mathbf{Y}}$ and $\mathbf{A}$ are independently distributed , from eq. (4) , $\underset{\sim}{\mathbf{Y}}^*$ and $\mathbf{A}^*$ are also independently distributed .

Let $\mathbf{\Omega} = (\omega_{ij})_{p\times p}$ is an orthogonal matrix in which first row is defined by

$$\omega_{1j} = \frac{Y^*_j}{\sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^*}} \quad , \text{ j=1,2,.....,p} \qquad \rightarrow (8)$$

where , Y^*_j is j^{th} component of $\underset{\sim}{\mathbf{Y}}^*$.

Now define , $\underset{\sim}{\mathbf{U}} = \mathbf{\Omega}\underset{\sim}{\mathbf{Y}}^*$

$$\mathbf{B} = \mathbf{\Omega}\mathbf{A}^*\mathbf{\Omega}' \qquad \rightarrow (9)$$

The i^{th} component of $\underset{\sim}{\mathbf{U}}$ is given by

$$U_i = \sum_{j=1}^{p} \omega_{ij} Y_j^*$$

$$= \sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^{*}} \sum_{j=1}^{p} \omega_{ij}\omega_{lj} \qquad \text{[using (8)]}$$

$$= \begin{cases} \sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^{*}} & \text{if } i = 1 \\ 0 & \text{otherwise} \end{cases}$$

(Since , $\boldsymbol{\Omega}$ is orthogonal matrix)

.

Thus , $$\underset{\sim}{\mathbf{U}} = \begin{bmatrix} \sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^{*}} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \rightarrow (10)$$

From equation (7) ,

$$\frac{T^2}{m} = \underset{\sim}{\mathbf{Y}}^{*\prime}\mathbf{I}\mathbf{A}^{*\prime}\mathbf{I}\underset{\sim}{\mathbf{Y}}^{*}$$

$$= \underset{\sim}{\mathbf{Y}}^{*\prime}\boldsymbol{\Omega}'\boldsymbol{\Omega}\mathbf{A}^{*\prime}\boldsymbol{\Omega}'\boldsymbol{\Omega}\underset{\sim}{\mathbf{Y}}^{*} \qquad (\because \boldsymbol{\Omega} \text{ is orthogonal})$$

$$= \left(\boldsymbol{\Omega}\underset{\sim}{\mathbf{Y}}^{*}\right)' \left(\boldsymbol{\Omega}\mathbf{A}^{*}\boldsymbol{\Omega}'\right)^{-1} \left(\boldsymbol{\Omega}\underset{\sim}{\mathbf{Y}}^{*}\right) \qquad (\because \boldsymbol{\Omega}^{-1} = \boldsymbol{\Omega}')$$

$$= \underset{\sim}{\mathbf{U}}'\mathbf{B}^{-1}\underset{\sim}{\mathbf{U}} \qquad \text{[using (9)]}$$

$$= \left(\sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^{*}} \quad 0 \quad \cdots \quad 0\right) \begin{bmatrix} b^{11} & b^{12} & \cdots & b^{1p} \\ b^{21} & b^{22} & \cdots & b^{2p} \\ \vdots & \vdots & & \vdots \\ b^{p1} & b^{p2} & \cdots & b^{pp} \end{bmatrix} \begin{bmatrix} \sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^{*}} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$= b^{11}\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^{*} \qquad \rightarrow (11)$$

where , b^{11} is first diagonal element of $\mathbf{B}^{-1}$.
But we know that ,

$$b^{11} = \frac{1}{b_{11} - \underset{\sim}{\mathbf{b}_1}' \mathbf{B}_{22}^{-1} \underset{\sim}{\mathbf{b}_1}} \quad , \quad \text{where} \quad \mathbf{B} = \begin{bmatrix} b_{11} & \underset{\sim}{\mathbf{b}_1}' \\ \underset{\sim}{\mathbf{b}_1} & b_{22} \end{bmatrix}$$

Thus , from eq (11) ,

$$\frac{T^2}{m} = \frac{\underset{\sim}{\mathbf{Y}}^{*\prime} \underset{\sim}{\mathbf{Y}}^{*}}{b_{11.2.....p}} \qquad \rightarrow (12)$$

where , $b_{11.2.....p} = b_{11} - \underset{\sim}{\mathbf{b}_1}' \mathbf{B}_{22}^{-1} \underset{\sim}{\mathbf{b}_1}$.

Let us suppose that $\mathbf{\Omega}$ is fixed (given). Then, just as we show $\mathbf{A}^*$ is distributed as $\sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{Z}}_{\alpha}^{*} \underset{\sim}{\mathbf{Z}}_{\alpha}^{*\prime}$, we can show that $\mathbf{\Omega A}^* \mathbf{\Omega}'$ is distributed as $\sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{V}}_{\alpha} \underset{\sim}{\mathbf{V}}_{\alpha}'$, when $\underset{\sim}{\mathbf{V}}_{\alpha} = \mathbf{\Omega} \underset{\sim}{\mathbf{Z}}_{\alpha}^{*}$ and $\underset{\sim}{\mathbf{V}}_{\alpha}$'s are i.i.d $\sim N_p(\underset{\sim}{\mathbf{0}}, \mathbf{I})$.

Now, with little difficult , we may show that $b_{11.2.....p} = b_{11} - \underset{\sim}{\mathbf{b}_1}' \mathbf{B}_{22}^{-1} \underset{\sim}{\mathbf{b}_1}$ is conditionally distributed as $\sum_{\alpha=1}^{m-(p-1)} \omega_{\alpha}^2$,

where $each\ \omega_{\alpha}$ is $i.i.d \sim N(0,1)$.

Therefore, $\sum_{\alpha=1}^{m-(p-1)} \omega_{\alpha}^2 \sim \chi_{m-(p-1)}^2$.

More over, the conditional distribution of $b_{11.2.....p}$ does not dependent on $\mathbf{\Omega}$, we have $b_{11.2.....p}$ is unconditionally distributed as $\chi_{m-(p-1)}^2$.

Also, since $\underset{\sim}{\mathbf{Y}}^{*} \sim N_p(\underset{\sim}{\mathbf{v}}^{*}, \mathbf{I})$, $\underset{\sim}{\mathbf{Y}}^{*\prime} \underset{\sim}{\mathbf{Y}}^{*} = \sum_{i=1}^{p} Y_i^{*2}$

where , $Y_i^* \sim N(v_i^*, 1)$ and Y_i^*'s are independent .

Thus , $\underset{\sim}{\mathbf{Y}}^{*\prime} \underset{\sim}{\mathbf{Y}}^{*}$ has non-central χ^2-distribution with non-centrality

$$= \sum_{i=1}^{p} v_i^{*2} = \underset{\sim}{\mathbf{v}}^{*\prime} \underset{\sim}{\mathbf{v}}^{*} = \underset{\sim}{\mathbf{v}}' \mathbf{C}' \mathbf{C} \underset{\sim}{\mathbf{v}} \qquad \text{[from(5)]}$$

$$= \underset{\sim}{\mathbf{v}}' \mathbf{\Sigma}^{-1} \underset{\sim}{\mathbf{v}} \qquad \text{[from (3)]}$$

Thus , $\frac{T^2}{m}$ is distributed as the ratio of non-central χ_p^2 and an independent central $\chi_{m-(p-1)}^2$.

Thus , $\frac{T^2}{m}\left[\frac{m-(p-1)}{p}\right] \sim F_{(p,m-p+1)}$ (non-central) and non-centrality parameter $\underset{\sim}{\nu}'\Sigma^{-1}\underset{\sim}{\nu}$.

If $\underset{\sim}{\nu} = \mathbf{0}$ then, $\underset{\sim}{\nu}'\Sigma^{-1}\underset{\sim}{\nu} = \mathbf{0}$ and therefore in this case , the distribution is central . $F_{(p,m-p+1)}$,the distribution of T^2 is called T^2-distribution with 'm' degrees of freedom .

COMPARISION OF SEVERAL MULTIVARIATE POPULATION MEANS

MANOVA for one way classification:

Suppose we have 'g' populations , each is distributed multivariate normal with mean vectors $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, ..., \underset{\sim}{\mu}_g$ respectively . Let us suppose that all populations have the same covariance matrix Σ. Thus, we have the 'g' populations.

$$\Pi_1 \sim N_p(\underset{\sim}{\mu}_1, \Sigma)$$
$$\Pi_2 \sim N_p(\underset{\sim}{\mu}_2, \Sigma)$$
$$\vdots$$
$$\Pi_g \sim N_p(\underset{\sim}{\mu}_g, \Sigma)$$

Now , we have a sample of size 'n_i' from i^{th} population Π_i . Thus, we have 'g' samples from the 'g' populations as follows :

$$\text{Population } \Pi_1 \ : \ \underset{\sim}{X}_{11}, \underset{\sim}{X}_{12}, ..., \underset{\sim}{X}_{1n_1}$$
$$\text{Population } \Pi_2 \ : \ \underset{\sim}{X}_{21}, \underset{\sim}{X}_{22}, ..., \underset{\sim}{X}_{2n_1}$$
$$\vdots$$
$$\text{Population } \Pi_g : \underset{\sim}{X}_{g1}, \underset{\sim}{X}_{g2}, ..., \underset{\sim}{X}_{gn_1}$$

Using the above random samples , MANOVA is used to investigate whether the population mean vectors are same and if not , which mean components differ significantly. Thus , the null hypothesis is

$$H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = ... = \underset{\sim}{\mu}_g \qquad \rightarrow (1)$$

ASSUMPTIONS CONCERNING THE STRUCTURE OF THE DATA :-

(1) $\underset{\sim}{X}_{i1}, \underset{\sim}{X}_{i2}, ..., \underset{\sim}{X}_{in_i}$ is a random sample of size n_i form a population with mean $\underset{\sim}{\mu}_i$,i=1,2,…..g.The random samples from different populations are independent.

(2) All populations have a common covariance matrix Σ .

(3) Each population is a multivariate normal. Condition (3) can be relaxed by appealing to the central limit theorem , when the sample sizes n_i's are large . If the mean vector of i^{th} population is written as

$$\underset{\sim}{\boldsymbol{\mu}}_{\mathbf{i}} = \underset{\sim}{\boldsymbol{\mu}} + \underset{\sim}{\boldsymbol{\tau}}_{\mathbf{i}} \qquad \rightarrow (2)$$

Here, $\underset{\sim}{\boldsymbol{\mu}}$ is the overall mean vector of all population and $\underset{\sim}{\boldsymbol{\tau}}_{\mathbf{i}}$ is a component due to the specific population , then the null hypothesis (1) can be written as

$$H_0 : \underset{\sim}{\boldsymbol{\tau}}_{\mathbf{1}} = \underset{\sim}{\boldsymbol{\tau}}_{\mathbf{2}} = \ldots\ldots = \underset{\sim}{\boldsymbol{\tau}}_{\mathbf{g}} = 0 \qquad \rightarrow (3)$$

The response $\underset{\sim}{\mathbf{X}}_{\mathbf{ij}}$, distributed as $N_p(\underset{\sim}{\boldsymbol{\mu}} + \underset{\sim}{\boldsymbol{\tau}}_{\mathbf{i}}, \boldsymbol{\Sigma})$, can be expressed in the suggestive form,

$$\underset{\sim}{\mathbf{X}}_{\mathbf{ij}} = \underset{\sim}{\boldsymbol{\mu}} + \underset{\sim}{\boldsymbol{\tau}}_{\mathbf{i}} + \underset{\sim}{\boldsymbol{\varepsilon}}_{\mathbf{ij}} \qquad \rightarrow (4)$$

$$\begin{pmatrix}\text{overall}\\ \text{mean}\end{pmatrix} \begin{pmatrix}\text{treatment}\\ \text{effect}\end{pmatrix} \begin{pmatrix}\text{random}\\ \text{error}\end{pmatrix}$$

i = 1,2,…..,g & j = 1,2,….., n_i .

where , $\underset{\sim}{\boldsymbol{\varepsilon}}_{\mathbf{ij}} \sim N(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma})$ are independent random variables. (4) is called as MANOVA model for comparing of population mean vectors. Here $\underset{\sim}{\boldsymbol{\mu}}$ is overall mean vector and $\underset{\sim}{\boldsymbol{\tau}}_{\mathbf{i}}$ represents the i^{th} treatment effect with

$$\sum_{i=1}^{g} n_i \underset{\sim}{\boldsymbol{\tau}}_{\mathbf{i}} = \underset{\sim}{\mathbf{0}} \qquad \rightarrow (5)$$

A vector of observations may be decomposed as suggested by model (4) . Thus,

$$\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} = \underset{\sim}{\bar{\mathbf{x}}} + (\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}}) + (\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}}) \qquad \rightarrow (6)$$

$$\begin{pmatrix}\text{observation}\end{pmatrix} \begin{pmatrix}\text{overall}\\ \text{sample}\\ \text{mean } \underset{\sim}{\boldsymbol{\mu}}\end{pmatrix} \begin{pmatrix}\text{estimated}\\ \text{treatment}\\ \text{effect } \underset{\sim}{\boldsymbol{\tau}}_{\mathbf{i}}\end{pmatrix} \begin{pmatrix}\text{residual}\\ \hat{\boldsymbol{\varepsilon}}_{\mathbf{ij}}\end{pmatrix}$$

When $\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}}$ = mean of i^{th} sample $\underset{\sim}{\mathbf{x}}_{\mathbf{i1}}, \underset{\sim}{\mathbf{x}}_{\mathbf{i2}}, \ldots\ldots, \underset{\sim}{\mathbf{x}}_{\mathbf{in_i}}$

$$\underset{\sim}{\bar{\mathbf{x}}} = \frac{1}{g}\left(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{1}} + \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{2}} + \ldots\ldots + \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{g}}\right) \qquad \text{(general mean)}$$

From (6), we may write the cross product,

$$\left(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}\right)' = \left(\left(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}}\right) + \left(\underset{\sim}{\bar{\mathbf{x}}}_{i} - \underset{\sim}{\bar{\mathbf{x}}}\right)\right)\left(\left(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}}\right) + \left(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}}\right)\right)'$$

$$= \left(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}}\right)' + \left(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}}\right)\left(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}}\right)'$$

$$+ \left(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}}\right)\left(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}}\right)' + \left(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}}\right)'$$

Summing the cross product over i and j , we get

$$\sum_{\mathbf{i=1}}^{\mathbf{g}}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}})(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}})' = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}})(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}})'$$
$$+\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}})(\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}}-\overline{\underset{\sim}{\mathbf{x}}})'$$
$$+\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}}-\overline{\underset{\sim}{\mathbf{x}}})(\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}})'$$
$$+\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}}-\overline{\underset{\sim}{\mathbf{x}}})(\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}}-\overline{\underset{\sim}{\mathbf{x}}})' \qquad \rightarrow (7)$$

But, since $\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}})=\underset{\sim}{\mathbf{0}}$., Eq (7) becomes ,

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}})(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}})' = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}})(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}})'$$
$$+\sum_{i=1}^{g}n_i(\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}}-\overline{\underset{\sim}{\mathbf{x}}})(\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}}-\overline{\underset{\sim}{\mathbf{x}}})' \qquad \rightarrow (8)$$

$$\Rightarrow \begin{pmatrix}\text{total(corrected)}\\ \text{sum of square}\\ \text{\& cross products}\end{pmatrix} = \begin{pmatrix}\text{residual(within)}\\ \text{sum of squares}\\ \text{\& cross products}\end{pmatrix} + \begin{pmatrix}\text{treatment(between)}\\ \text{sum of squares}\\ \text{\& cross products}\end{pmatrix}$$

That is (8) may be written as

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}})(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}})' = W + B \qquad \rightarrow (9)$$

where , $W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}})(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}}-\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}})'$

$$= (n_1-1)S_1 + (n_2-1)S_2 + \ldots\ldots + (n_g-1)S_g$$

where , S_i is sample covariance matrix of i^{th} sample .

and $B = \sum_{i=1}^{g} n_i(\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}}-\overline{\underset{\sim}{\mathbf{x}}})(\overline{\underset{\sim}{\mathbf{x}}}_{\mathbf{i}}-\overline{\underset{\sim}{\mathbf{x}}})'$.

Now, we summarise the calculations leading to the test statistic in a MANOVA table . MANOVA table for comparing population mean vectors :-

Source of variation	Matrix of sum of squares & cross product	Degrees of freedom

Treatments	$B = \sum_{i=1}^{g} n_i (\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})'$	g-1
Residual(error)	$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i)(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i)'$	n-g
Total (correlated for the mean)	$B + W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}})'$	n-1

Now one of the test statistic for testing (3) involves generalized variances and is given by

$$\Lambda^* = \frac{|W|}{|B+W|} \qquad \rightarrow (10)$$

The quantity Λ^* is called Wilk's lamda and related to likelihood ratio criterion. The exact distribution of Λ^* can be derived for the special cases listed in the following table .

Distribution of Wilk's lamda , Λ^* :-

No . of variables	No .of groups	Sampling distribution for multivariate normal data
p - 1	g ≥ 2	$\left[\frac{\sum_{i=1}^{g} n_i - g}{g-1}\right]\left[\frac{1-\Lambda^*}{\Lambda^*}\right] \sim F_{\left(g-1, \sum_{i=1}^{g} n_i - g\right)}$
p = 2	g ≥ 2	$\left[\frac{\sum_{i=1}^{g} n_i - g - 1}{g-1}\right]\left[\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right] \sim F_{\left(2(g-1), 2\sum_{i=1}^{g} n_i - g - 1\right)}$
p ≥ 1	g = 2	$\left[\frac{\sum_{i=1}^{g} n_i - p - 1}{p}\right]\left[\frac{1-\Lambda^*}{\Lambda^*}\right] \sim F_{\left(p, \sum_{i=1}^{g} n_i - p - 1\right)}$

$p \geq 1$	$g = 3$	$\left[\frac{\sum_{i=1}^{g} n_i - p - 2}{p}\right]\left[\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right] \sim F_{\left(2p,\,2(\sum_{i=1}^{g} n_i - p - 2)\right)}$

Bartlett has shown that if H_o is true and $\sum_{i=1}^{g} n_i = n$ is large,

$$-\left[n - 1 - \frac{(p+g)}{2}\right]\ln \Lambda^* = -\left[n - 1 - \frac{(p+g)}{2}\right]\ln \frac{|W|}{|B+W|}$$

has approximately a χ^2 - distribution with p(g-1) d.f. consequently.

Principle Component Analysis

Introduction :

Suppose $X_1, X_2, ..., X_p$ are the given random variables. Then, principle component analysis (P.C.A) is concerned with explaining the variance-covariance structure of the variables through a few standardized linear combinations (SLC) of the original variables (we call a linear combination $l_1X_1 + l_2X_2 + ... + l_pX_p$ as an SLC if $\sum_i l_i^2 = 1$).

Algebraically principal components (PCs) are particular standard linear combinations (SLCs) of the components of the original pattern and geometrically, these LCs represent the selection of new coordinate system obtained by rotating the original system with X_1, X_2,..., X_p as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious (avoiding of excess) description of the covariance structure . As we shall see , principle components depend solely on the covariance matrix (or the correlation matrix) of the random variables $X_1, X_2, ..., X_p$. Their development does not require a multivariate normal assumption.

The general objections of P.C.A are

(i) Data-reduction and

(ii) Interpretation.

Although p components required to reproduce the total system variability, often much of this variability can be accounted by a small number 'k'(<p) of the principal components. If there is almost as much information in the k components as there is in the original 'p' variables, then the 'k' principal components replace the original 'p' components of the pattern. And the original data set consisting of n measurements on p-component pattern is reduced to one consisting of n measurements on k-principal component pattern. In other

words, PCA reduces the dimensionality of the given data, losing as little information as possible. This technique was developed by Hotelling(1933).

An analysis of principle components often reveals relationships that were not previously suspected and there by allows interpretations that would not ordinarily result. In other words, the key problem is the interpretation of the principle components.

PCs may be inputs to a multiple regression analysis or cluster analysis. Moreover , (scaled) principle components are one factoring of the covariance matrix for the factor analysis model.

An example:-

Suppose we consider a sample of ***n*** students and they are asked to write five papers mechanics (X_1), vectors (X_2), algebra (X_3), analysis(X_4) and statistics (X_5). The examination in the first two papers is conducted in the closed book system , where as in the remaining three papers in the open book system.

Thus , we have totally '5n' observations so that ***n*** observations on each paper . One question which can be asked concerning this data is how the results on the five different papers should be combined to produce an overall scare various answers are possible. One obvious answer would be to use the overall mean that is the linear combination $(X_1 + X_2 + X_3 + X_4 + X_5)/5$. But , can one do better than this? . This is one of the questions that principle component analysis seeks to answer .

Definition Of Principle Component :-

Definition:-

If $\underset{\sim}{X}$ is a random vector with mean $\underset{\sim}{\mu}$ and variance – covariance matrix Σ, then the principle component transformation is the transformation.

$$\underset{\sim}{X} \to \underset{\sim}{Y} = \Omega' (\underset{\sim}{X} - \underset{\sim}{\mu}) \qquad \to (1)$$

where , Ω is orthogonal matrix , such that

$\Omega'\Sigma\Omega=\Lambda=$ diag$(\lambda_1,\lambda_2,.....,\lambda_p)$, $\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_p > 0$

The strict positivity of the eigen values λ_i is guaranteed if, Σ is positive definite. The i^{th} principle component of $\underset{\sim}{X}$ may be defined as the i^{th} element of the vector $\underset{\sim}{Y}$, namely as $Y_i = \underset{\sim}{\omega}_i'(\underset{\sim}{x}-\underset{\sim}{\mu})$, where $\underset{\sim}{\omega}_i$ is the i^{th} column of Ω and may be called the i^{th} vector of principle components leadings.

An Alternative Definitions:-

If $\underset{\sim}{X}$ is a px1 random vector with mean $\underset{\sim}{\mu}$ and variance-covariance matrix Σ, then the first principle component of $\underset{\sim}{X}$ is defined by the Standardized Linear Combination (SLC) of $\underset{\sim}{X}$.

$$Y_1 = \underset{\sim}{\beta}_1'\underset{\sim}{X} \qquad (\text{ where, } \beta_1'\beta_1 = 1)$$

such that $V(Y_1)$ is larger than the variance of any other SLC $\underset{\sim}{\alpha}'\underset{\sim}{X}$.

i.e., $V(Y_1) \geq V(\underset{\sim}{\alpha}',\underset{\sim}{X})$ for an arbitrary '$\underset{\sim}{\alpha}$' ($\underset{\sim}{\alpha}'\underset{\sim}{\alpha}=1$). In otherwords Y_1 has the largest variance among all SLC's of $\underset{\sim}{X}$.

Principle Component Definition :-

In general the k^{th} principle component of $\underset{\sim}{X}$ is defined by the SLC of $\underset{\sim}{X}$, $Y_k = \underset{\sim}{\beta}_k'\underset{\sim}{X}$ (where, $\boldsymbol{\beta}_k'\boldsymbol{\beta}_k=1$), which is uncorrelated with first k-1 principle component and $V(Y_k) \geq V(Y_i)$, for i=1,2,……,k-1 .

Derivation Of The Principle Components:-

Suppose $\underset{\sim}{X}$ is px1 random vector with mean vector $\underset{\sim}{\mu}$ and covariance-matrix Σ i.e., $\underset{\sim}{X} \sim (\underset{\sim}{\mu},\Sigma)$, then by definition, the first principle component is the SLC of $\underset{\sim}{X}$ which has largest variance among all SLC's of $\underset{\sim}{X}$. Thus we should seek a LC of $\underset{\sim}{X}$ viz.,

$$Y = \underset{\sim}{\omega}'\underset{\sim}{X} \qquad \rightarrow (1)$$

with largest variance,

$$V(Y) = \underset{\sim}{\omega}' \, \mathbf{V}(\underset{\sim}{X}) \, \underset{\sim}{\omega} = \underset{\sim}{\omega}' \, \mathbf{\Sigma} \, \underset{\sim}{\omega} \qquad \rightarrow (2)$$

such that $\underset{\sim}{\omega}'\underset{\sim}{\omega} = 1$.

Thus , we have to maximize (2) subject to the condition

$$\underset{\sim}{\omega}'\underset{\sim}{\omega} = 1 \qquad \rightarrow (3)$$

which is equivalent to maximizing the function ,

$$\phi(\underset{\sim}{\omega},\lambda) = \underset{\sim}{\omega}'\mathbf{\Sigma}\underset{\sim}{\omega} - \lambda(\underset{\sim}{\omega}'\underset{\sim}{\omega}-1) \qquad \rightarrow (4)$$

w.r.t $\underset{\sim}{\omega}$ and λ , where 'λ' is a Lagranges multiplier. This implies to solve the equations , $\frac{\partial \phi}{\partial \underset{\sim}{\omega}} = \underset{\sim}{0} \quad \Rightarrow \quad \Sigma\underset{\sim}{\omega} = \lambda\underset{\sim}{\omega}$

$$\text{i.e.,} \quad (\Sigma-\lambda I)\underset{\sim}{\omega} = \underset{\sim}{0} \qquad \rightarrow (5)$$

$$\frac{\partial \phi}{\partial \lambda} = 0 \quad \Rightarrow \quad \underset{\sim}{\omega}'\underset{\sim}{\omega} = 1 \qquad \rightarrow (6)$$

Using (5) & (6) , from (2) , we get , $V(Y) = \lambda\underset{\sim}{\omega}'\underset{\sim}{\omega} = \lambda \qquad \rightarrow (7)$

From (5) , to have a non-zero solution for $\underset{\sim}{\omega}$, we must have ,

$$|\Sigma-\lambda I| = 0 \qquad \rightarrow (8)$$

We know that (8) is a characteristic equation , and 'λ' is a latent root and From (5) , $\underset{\sim}{\omega}$ is the corresponding latent vector of the equation . But, we know that , solving (8) for 'λ' gives p-latent roots (positive) ,

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots\ldots\ldots\ldots \geq \lambda_p \geq 0 \qquad \rightarrow (9)$$

With the corresponding latent vectors , $\omega_1,\omega_2,\ldots\ldots\ldots\ldots,\omega_p$ respectively ,

i.e., we have from (5) , $\Sigma\underset{\sim}{\omega}_i = \lambda_i\underset{\sim}{\omega}_i$, i= 1,2,……p $\qquad \rightarrow (9.a)$

Since λ_1 is the largest latent root among all latent roots and $\underset{\sim}{\omega}_1$ is the corresponding latent vector .

From (1) and (7) , $Y_1 = \underset{\sim}{\omega}_1'\underset{\sim}{X}$, is the first principle component with variance, $V(Y_1) = \lambda_1$.

Let us denote the first principle component by Y_1 . Now , $Y_1 = \underset{\sim}{\omega}_1'\underset{\sim}{X} \qquad \rightarrow (10)$

$$V(Y_1) = \lambda_1 \qquad \rightarrow (11)$$

Now , let us show that for $2 \leq k \leq p$, $Y_K = \underset{\sim}{\omega}_K'\underset{\sim}{X} \qquad \rightarrow (12)$

is the K^{th} principle component with variance , $V(Y_K) = \lambda_K \qquad \rightarrow (13)$

By definition , Y_K should uncorrelated with $Y_1,Y_2,\ldots\ldots\ldots\ldots Y_{K-1}$, which can be

easily verified as follows (for J = 1,2,……..k-1) .

$$\text{Cov}(Y_k, Y_i) = \text{Cov}(\boldsymbol{\omega}_k' \mathbf{X}, \boldsymbol{\omega}_i' \mathbf{X})$$

$$= \boldsymbol{\omega}_k' \text{ Cov}(\mathbf{X}, \mathbf{X}') \boldsymbol{\omega}_i$$

$$= \boldsymbol{\omega}_k' \boldsymbol{\Sigma} \boldsymbol{\omega}_i$$

$$= \lambda_i \boldsymbol{\omega}_k' \boldsymbol{\omega}_i \qquad \text{[From(5)]}$$

$$= 0 \quad (\because \underset{\sim}{\omega}_K \ \& \ \underset{\sim}{\omega}_J \text{ are orthogonal vectors})$$

Also by definition, the k^{th} PC Y_k has largest variance than $Y_{k+1},\ldots,Y_p$ which can also be verifying from (13).

$$\lambda_K \geq \lambda_{K+1} \geq \ldots\ldots\ldots\ldots \geq \lambda_P \geq 0$$

$$\Rightarrow V(Y_K) \geq V(Y_{K+1}) \geq \ldots\ldots\ldots\ldots \geq V(Y_P) \geq 0 .$$

Hence the proof .

Remark :-

The above result may be asked as no standard linear combination (SLC) of $\underset{\sim}{X}$ has a variance larger than λ_1 ,the variance of first principle combination.

From the above result , we may say that construction (derivation) of principle components of a given random vector $\underset{\sim}{X}$ is equivalent to the problem of the construction (derivation) of the latent roots and latent vectors of the variance-covariance matrix Σ of $\underset{\sim}{X}$ in case of known Σ .

Note:-

(1) If Σ is not known , we may construct the principle component of the random vector $\underset{\sim}{X}$ based on the sample variance – covariance matrix or sample correlation matrix .

(2) If the population correlation matrix ' ρ ' is given , we may use it in place of Σ to construct the principle components .

(3) The principle components of the random vector $\underset{\sim}{X}$ derived from population (sample) correlation matrix are different from the principle components derived from population (sample) covariance matrix.

Properties Of Principle Components :-

(1) Sum of the variances of all p.c's equal to the trace of Σ.

Proof :- Let $Y_1, Y_2, \ldots\ldots\ldots\ldots, Y_p$ are the p.c's obtained from random variable $\underset{\sim}{X}$.

Let us denote, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots\ldots\ldots\ldots, \lambda_p)_{p\times p}$

$$\Omega = (\underset{\sim}{\omega}_1, \underset{\sim}{\omega}_2, \ldots\ldots\ldots\ldots, \underset{\sim}{\omega}_p)_{p\times p}$$

where, λ_i's are the latent roots and $\underset{\sim}{\omega}_i$'s are the latent vectors of the covariance matrix Σ of the random variable $\underset{\sim}{X}$. Then, we have,

$$\Omega'\Sigma\Omega = \Lambda .$$

$$\Rightarrow \text{Tr}(\Lambda) = \text{Tr}(\Omega'\Sigma\Omega)$$

$$\Rightarrow \lambda_1+\lambda_2+\ldots\ldots\ldots\ldots+\lambda_p = \text{Tr}(\Sigma\Omega\Omega')$$

$$\Rightarrow V(Y_1)+V(Y_2)+\ldots\ldots\ldots\ldots+V(Y_p) = \text{Tr}(\Sigma I)$$

$$= \text{Tr}(\Sigma)$$

$$= V(X_1)+V(X_2)+\ldots\ldots\ldots\ldots+V(X_p)$$

$$(\lambda_i = V(Y_i)).$$

(2) Product of the variance of p.c's is equal to the determinant of Σ i.e., $|\Sigma|$ (or generalized variance).

Proof :- We have $\Omega'\Sigma\Omega = \Lambda$

$$\Rightarrow |\Lambda| = |\Omega'\Sigma\Omega|$$

$$\Rightarrow \lambda_1\lambda_2\ldots\ldots\ldots\ldots\lambda_p = |\Sigma\Omega\Omega'|$$

$$\Rightarrow V(Y_1)V(Y_2)\ldots\ldots\ldots\ldots V(Y_p) = |\Sigma I|$$

$$= |\Sigma|$$

Hence the proof.

(3) The sum of the first K eigen values divided by the sum of all eigen values,

$$\frac{\lambda_1+\lambda_2+\ldots\ldots\ldots\ldots+\lambda_k}{\lambda_1+\lambda_2+\ldots\ldots\ldots\ldots+\lambda_p} = \frac{\lambda_1+\lambda_2+\ldots\ldots\ldots\ldots+\lambda_k}{\text{Tr}(\Sigma)}$$

represents the 'Proportion of total variation' explained by the first K principle components.

(4) The principle components of a random vector are not scale invariant. It is one disadvantage of principle component analysis.

Theorem :-

An orthogonal transformation $\underset{\sim}{Y} = C\underset{\sim}{X}$ of a random vector $\underset{\sim}{X}$ leaves invariant the generalized variance and the sum of the variance of the components .

Proof :- We have given $\underset{\sim}{X}$ is the original random vector and $\underset{\sim}{Y}$ is the transformed random variable using the orthogonal matrix C . Now, we have to show that , $\left|\text{cov}\left(\underset{\sim}{X},\underset{\sim}{X}'\right)\right| = \left|\text{cov}\left(\underset{\sim}{Y},\underset{\sim}{Y}'\right)\right|$, where $|\ |$ is determinant and $\sum_{i=1}^{p} V(X_i) = \sum_{i=1}^{p} V(Y_i)$

since , 'C' is orthogonal we have, $C'C = CC' = I \quad \rightarrow (1)$

Now , $\text{cov}\,(Y,Y') = \text{cov}\,(C\underset{\sim}{X},(C\underset{\sim}{X})')$

$= C\,\text{cov}\,(\underset{\sim}{X},\underset{\sim}{X}')C'$

$= C\Sigma C'$

$\Rightarrow \left|\text{cov}(\underset{\sim}{Y},\underset{\sim}{Y}')\right| = \left|C\Sigma C'\right|$

$= |C||\Sigma||C'|$

$= |\Sigma||CC'|$

$= |\Sigma||I| \qquad (\because \text{ from (1)})$

$= |\Sigma|$

$= \left|\text{cov}\left(\underset{\sim}{X},\underset{\sim}{X}'\right)\right|$

$\Rightarrow$ generalized variance of $\underset{\sim}{Y}$ = generalized variance of $\underset{\sim}{X}$.

We have , $\sum_{i=1}^{p} V(X_i) = \text{Tr}\,(\Sigma)$

$= \text{Tr}\,(\Sigma I)$

$= \text{Tr}\,(\Sigma\, CC') \qquad (\because \text{ from (1)})$

$= \text{Tr}\,(C\Sigma C')$

$= \sum_{i=1}^{p} V(Y_i) \qquad (\because C\Sigma C' \text{ is covariance matrix of } \underset{\sim}{Y}) \Rightarrow$

Sum of the variances of original variables (total population variance)

= Sum of variances of principle components .

$= \lambda_1 + \lambda_2 + \ldots\ldots\ldots\ldots\ldots\ldots + \lambda_p$.

Note :-

The above theorem may be stated as follows. The generalized variance of the vector of principle components is the generalized variance of the original vector and the sum of the variances of the principle components is the sum of the variances of the original variates.

Results :-

If $\underset{\sim}{X}$ is a random vector with covariance matrix Σ and $Y_i = \underset{\sim}{\omega}_i' \underset{\sim}{X}$ is the i^{th} principle component of the random vector $\underset{\sim}{X} = (X_1, X_2, \ldots\ldots\ldots\ldots, X_p)'$, then the correlation coefficient between i^{th} principle component and J^{th} original variable (that correlation coefficient between Y_i and X_J) is given by

$$\rho_{Y_i, X_J} = \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{JJ}}} \omega_{iJ} \qquad i, J = 1, 2, \ldots\ldots\ldots, p \quad , \quad \text{where } \sigma_{JJ} = V(X_J)$$

λ_i is i^{th} largest root of Σ and ω_{iJ} is J^{th} component of $\underset{\sim}{\omega}_i$, when $\underset{\sim}{\omega}_i$ is the latent vector of Σ corresponding to λ_i.

Proof :- Denote $\underset{\sim}{l}_J = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ J \\ 0 \\ 0 \end{pmatrix} \rightarrow J^{th}$ position

Now, $X_J = \underset{\sim}{l}_J' \underset{\sim}{X}$. Also we have given, $Y_i = \underset{\sim}{\omega}_i' \underset{\sim}{X} \qquad \rightarrow (1)$

$$\rho_{Y_i, X_J} = \frac{\text{cov}\,(\underset{\sim}{\omega}_i' \underset{\sim}{X}\,,\, \underset{\sim}{l}_J' \underset{\sim}{X})}{\sqrt{\lambda_i \sigma_{JJ}}} \quad \frac{\text{cov}\,(Y_i, X_J)}{\sqrt{V(Y_i) V(X_J)}}$$

$$= \frac{\text{cov}\,(\underset{\sim}{\omega}_i' \underset{\sim}{X}\,,\, \underset{\sim}{l}_J' \underset{\sim}{X})}{\sqrt{\lambda_i \sigma_{JJ}}}$$

($\because V(Y_i) = \lambda_i$, the i^{th} larger latent root of Σ, using (1) & σ_{JJ} is J^{th} diagonal element of Σ).

$$= \frac{\underset{\sim}{\omega}_i' \,\text{cov}\,(\underset{\sim}{X}\,,\, \underset{\sim}{X}')\, \underset{\sim}{l}_J}{\sqrt{\lambda_i}\sqrt{\sigma_{JJ}}}$$

$$= \frac{\underset{\sim}{\omega}_i' \,\Sigma\, \underset{\sim}{l}_J}{\sqrt{\lambda_i}\sqrt{\sigma_{JJ}}} \qquad \rightarrow (2)$$

($\because \Sigma$ is covariance matrix of $\underset{\sim}{X}$)

Since $\underset{\sim}{\omega}_i$ is the latent vector of Σ corresponding to latent root λ_i, we have

$$\Sigma \underset{\sim}{\omega}_i = \lambda_i \underset{\sim}{\omega}_i$$

$$\Rightarrow \underset{\sim}{\omega}_i' \Sigma = \lambda_i \underset{\sim}{\omega}_i' \quad \text{(Taking transpose \& } \Sigma = \Sigma' \text{)} \qquad \rightarrow (3)$$

Using (3) in (2), we get ,

$$\rho_{Y_i,X_J} = \frac{\lambda_i \underset{\sim}{\omega}_i' \, l_J}{\sqrt{\lambda_i}\sqrt{\sigma_{JJ}}}$$

$$= \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{JJ}}} \; (\omega_{i1}0+\omega_{i2}0+.....+\omega_{iJ}1+\omega_{iJ+1}0+.....+0)$$

$$= \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{JJ}}} \; \omega_{iJ} \qquad \text{for } i,J = 1,2,\ldots\ldots\ldots,p$$

Hence the proof .

Principle Components Obtained From Standardized Variables :-

Suppose $\underset{\sim}{X}$ is a random vector with mean , $\underset{\sim}{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ : \\ : \\ \mu_p \end{pmatrix}$

and covariance matrix , $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & .... & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & .... & \sigma_{2p} \\ : & & & \\ \sigma_{p1} & \sigma_{p2} & .... & \sigma_{pp} \end{pmatrix}$

Now , the standardized random vector of $\underset{\sim}{X}$ is given by

$$\underset{\sim}{Z} = \begin{pmatrix} \frac{X_1-\mu_1}{\sqrt{\sigma_{11}}} \\ \frac{X_2-\mu_2}{\sqrt{\sigma_{22}}} \\ : \\ : \\ \frac{X_p-\mu_p}{\sqrt{\sigma_{pp}}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & .. & .. & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & .. & .. & 0 \\ : & : & & & \\ : & : & & & \\ 0 & 0 & .. & .. & \frac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix} \begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ : \\ : \\ X_p-\mu_p \end{pmatrix}$$

$$= D\,(\underset{\sim}{X} - \underset{\sim}{\mu}) \qquad \rightarrow (1)$$

Where , $D = \text{diag}\left(\frac{1}{\sqrt{\sigma_{11}}}, \frac{1}{\sqrt{\sigma_{22}}}, \ldots\ldots\ldots\ldots, \frac{1}{\sqrt{\sigma_{pp}}}\right)$

Now , $\text{cov}(\underset{\sim}{Z},\underset{\sim}{Z}') = D \; \text{cov}(\underset{\sim}{X}-\underset{\sim}{\mu}, (\underset{\sim}{X}-\underset{\sim}{\mu})') \; D'$

$$= D\Sigma D \qquad (\text{since } D = D' \;\&\; \text{cov}(\underset{\sim}{X},\underset{\sim}{X}') = \Sigma) \qquad \rightarrow (2)$$

But we verify that,

$$D\Sigma D = \begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \dots & \dots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \dots & \dots & 0 \\ \vdots & \vdots & & & \\ \vdots & \vdots & & & \\ 0 & 0 & \dots & \dots & \frac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \dots & \sigma_{2p} \\ \vdots & \vdots & & & \\ \vdots & \vdots & & & \\ \sigma_{p1} & \sigma_{p2} & \dots & \dots & \sigma_{pp} \end{pmatrix}$$

$$\begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \dots & \dots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \dots & \dots & 0 \\ \vdots & \vdots & & & \\ \vdots & \vdots & & & \\ 0 & 0 & \dots & \dots & \frac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} & \dots & \dots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}\sigma_{pp}}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_{11}\sigma_{22}}} & 1 & \dots & \dots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}\sigma_{pp}}} \\ \vdots & \vdots & & & \\ \vdots & \vdots & & & \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{11}\sigma_{pp}}} & \dots & \dots & \dots & 1 \end{pmatrix}$$

$= \rho$, the correlation matrix $\underset{\sim}{X}$. $\rightarrow (3)$

From (2) & (3) , $\text{cov}(\underset{\sim}{Z},\underset{\sim}{Z}') = \rho$

Thus , the covariance matrix of standardized random vector $\underset{\sim}{Z}$ is nothing but the correlation matrix of the original random vector $\underset{\sim}{X}$.

Now , the principle component of $\underset{\sim}{Z}$ may be obtained from the eigen vectors of the correlation matrix ρ of $\underset{\sim}{X}$. The i^{th} principle component of the standardized variables $Z_1, Z_2, \dots\dots\dots\dots, Z_p$ with $\text{cov}(\underset{\sim}{Z},\underset{\sim}{Z}') = \rho = \text{cov}(\underset{\sim}{X},\underset{\sim}{X}')$ is given by $Y_i = \underset{\sim}{\omega}_i' \underset{\sim}{z} = \underset{\sim}{\omega}_i' D (\underset{\sim}{X}-\underset{\sim}{\mu})$, i=1,2,…….p ($\because$ from (1))

Moreover , $\sum_{i=1}^{p} var(Y_i) = \sum_{i=1}^{p} \lambda_i = Tr(\rho)$

$$= \sum_{i=1}^{p} V(Z_i)$$

$$= p \qquad (\because V(Z_i) = 1)$$

and $cov(Y_i, Z_J) = \omega_{iJ}\sqrt{\lambda_i}$ where , i,J= 1,2,……..,p.

Here , $(\lambda_i, \underset{\sim}{\omega}_i)$, i=1,2,……,p are eigen value – eigen vector pairs of ' ρ ' with $\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_p$.

Note :-

The principle components discussed so far are called as population principle components . The computation of principle components is possible only when Σ or ρ is known , which is a rare practical situation . Therefore , we have to estimate the population principle components from the given sample . The estimation of the population principle components is equivalent to finding the eigen values and the eigen vectors of the sample covariance matrix S obtained from the sample . Therefore estimated population principle components are also called as Sample Principle Components. Sometimes , we may obtain the sample principle components from the sample correlation matrix R. But, the sample principle components obtained from S are different from those obtained from R . .

Result :-

If $S = (s_{iJ})_{p\times p}$ is the sample covariance matrix of the given sample $\underset{\sim}{x}_1, \underset{\sim}{x}_2, ..., \underset{\sim}{x}_n$ from a multivariate population and $(\hat{\lambda}_1, \hat{\underset{\sim}{\omega}}_1)$ $(\hat{\lambda}_2, \hat{\underset{\sim}{\omega}}_2)$ …… $(\hat{\lambda}_p, \hat{\underset{\sim}{\omega}}_p)$ are the eigen value – eigen vectors pairs of S , then the i^{th} sample principle component is given by

Consider a pattern (random vector)

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ X_p \end{pmatrix} \in R^p \qquad (1)$$

with Covariance matrix $\mathbf{\Sigma}$ given by

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & . & . & . & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & . & . & . & \sigma_{2p} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ \sigma_{p1} & \sigma_{p2} & . & . & . & \sigma_{pp} \end{bmatrix} \qquad (2)$$

where $\sigma_{ij} = Cov(X_i, X_j)$

FACTOR ANALYSIS

INTRODUCTION:-

Factor analysis is a mathematical model which attempts to explain the correlation between a large set of variables in terms of a small number of underlying unobservable factors. In other words , the essential purpose of factor analysis is to describe , if possible , the covariance relationships among many variables in terms of a few underlying but unobservable , random quantities called factors. Basically, the factor model is motivated by the following argument. Suppose variables can be grouped by their correlations . That is all variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. It is conceivable that each group of variables represents a single underlying construct, or factor , that is responsible for the observed correlations. Factor analysis was originally developed by psychologists interested in psychometric measurement.

Arguments over the psychological interpretations of several early studies and the lack of powerful computing facilities impelled its developments a statistical method. The advent of high speed computers has generated a renewed interst in the theoretical and computational aspects of factor analysis.most of the original techniques have been abandoned and early controversies resolved in the make of recent developments.It is till true that each application of the technique must be examined on its own merits to determine its success.

Factor analysis can be considered as an extension of principal component analysis. Both can be viewed as attempts to approximate the covariance matrix Σ .However, the approximation based

on the factor analysis model is more elaborate . The primary question is factor analysis is whether the data are consistent with a prescribed structure.

In order to get a feel for the subject we first describe a simple example .

Example 1 (Spearmon,1904): In children examinations performance in classics (x_1),French (x_2) and English (x_3) .It is found that the correlation matrix is given by

$$\begin{pmatrix} 1 & 0.83 & 0.78 \\ & 1 & 0.67 \\ & & 1 \end{pmatrix}$$

Although this matrix has full rank its dimantionality can be effectively reduced from p =3 to p=1by expressing the three variables as follows

$$\left.\begin{aligned} x_1 &= \lambda_1 f + u_1 \\ x_2 &= \lambda_2 f + u_2 \\ x_3 &= \lambda_3 f + u_3 \end{aligned}\right\} \qquad \textbf{--- (1)}$$

In these equations f is an underlying 'common factor' and λ_1, λ_2 and λ_3 are known as factor loadings.The terms u_1, u_2 and u_3 represent random disturbance terms. The common factor may be interpreted as 'general ability' (or 'intelligence') and u_i will have small variance is x_i is closely related to general ability. The variation is u_i consist of two parts which we shall not try to disentangle in practice.First,this variance represent the extent to which an individuals ability at classics , say, differs from his general ability and second it represents the fact that the examination is only an approximate measure of his ability in the subject.

The model defined in (1) can be generalized to include k >1 common factors.

ORTHOGONAL FACTOR MODEL:

The observable random vector $\underset{\sim}{\mathbf{x}}$ with p component has mean $\underset{\sim}{\boldsymbol{\mu}}$ and covariance matrix $\boldsymbol{\Sigma}$. The factor model postulates that $\underset{\sim}{\mathbf{x}}$ is linearly dependent upon a few unobservable random variables $F_1, \ldots, F_k$ called comman factors and p additional sources of variations $u_1, u_2, \cdots, u_p$ called random disturbances or error or specific factors . In particular, the factor analysis model is

$$\left.\begin{aligned} X_1 &= \mu_1 + \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1k}F_k + u_1 \\ X_2 &= \mu_2 + \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2k}F_k + u_2 \\ &\vdots \\ X_p &= \mu_p + \lambda_{p1}F_1 + \lambda_{p2}F_2 + \cdots + \lambda_{pk}F_k + u_p \end{aligned}\right\} \quad \cdots (1)$$

(or) in matrix notation.

$$\underset{\sim}{\mathbf{X}}_{(p\times 1)} = \underset{\sim}{\boldsymbol{\mu}}_{(p\times 1)} + \boldsymbol{\Lambda}_{(p\times k)}\underset{\sim}{\mathbf{F}}_{(k\times 1)} + \underset{\sim}{\mathbf{u}}_{(p\times 1)} \qquad \textbf{(2)}$$

Where $\underset{\sim}{\mathbf{X}}=\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}, \underset{\sim}{\boldsymbol{\mu}}=\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \underset{\sim}{\mathbf{F}}=\begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_k \end{pmatrix}$, $\underset{\sim}{\mathbf{u}}=\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}, \boldsymbol{\Lambda}=\begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2k} \\ \vdots & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pk} \end{pmatrix}$

The matrix $\boldsymbol{\Lambda}$ is called the matrix of factor loadings, where λ_{ij} is the loading of i^{th} variable (X_i) on j^{th} factor (F_j).Note that the i^{th} specific factor u_i is associated only with the i^{th} response X_i.

The p deviations $X_1-\mu_1, X_2-\mu_2, \cdots, X_p-\mu_p$ are expressed in terms of k+p random variables $F_1, F_2, \cdots, F_k, \mu_1, \mu_2, \cdots, \mu_p$ are unobservable.

From (1) ,it may be noted that each equation looks like a multiple regression equation but for one exception. The common factor in (1) $F_1, F_2, \cdots, F_k$ are unobservable where as in multiple regression equation the independent variables can be observed .This distinguishes the factor model from the multivariate regression model. With so many unobservable quantities (k+p) a direct verification of the factor model (1) from observations on $X_1, \ldots, X_p$ is hopeless. However, with some additional assumptions about the random vectors $\underset{\sim}{F}$ and $\underset{\sim}{\boldsymbol{\mu}}$, the model in(2) implies certain covariance relationships ,which can be checked.

We assume that

$$\left.\begin{aligned} &E(\underset{\sim}{\mathbf{F}})=\underset{\sim}{0}\ ,\ V(\underset{\sim}{\mathbf{F}})=E(\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{F}}')=I_{k\times k} \\ &E(\underset{\sim}{\mathbf{u}})=\underset{\sim}{0}\ ,\ V(\underset{\sim}{\mathbf{u}})=E(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}') \\ &\qquad\qquad =\boldsymbol{\psi}=\text{doag}(\psi_1,\psi_2,\cdots,\psi_p) \\ &\text{and } \text{cov}(\underset{\sim}{\mathbf{u}},\underset{\sim}{\mathbf{F}})=E(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{F}}')=0_{p\times k} \end{aligned}\right\} \quad \textbf{---- (3)}$$

The model (2) with the assumptions (3) is called the 'Orthogonal Factor model'
The assumption (3) implies the following implicit assumptions.

- All common factors are standardized to have variance 1 and uncorrelated with one another $(V(\underset{\sim}{\mathbf{F}})=I)$
- All specific factors (random disturbances) are have zero means and uncorrelated

$\left(V(\underset{\sim}{u})=\boldsymbol{\psi}=\text{doag}(\psi_1,\psi_2,\cdots,\psi_p)\right)$

- Common factor and specific factor are uncorrelated $\left(\text{cov}(\underset{\sim}{u},\underset{\sim}{F})=0\right)$.

The Orthogonal model with k common factors

$$\underset{\sim}{X}_{(p\times 1)}=\underset{\sim}{\mu}_{(p\times 1)}+\Lambda_{(p\times k)}\underset{\sim}{F}_{(k\times 1)}+\underset{\sim}{u}_{(p\times 1)} \qquad \text{---- (4)}$$

Where X_i = i^{th} response variable

μ_i = mean of X_i

λ_{ij} = loading of X_i on F_j

F_j = j^{th} common factor

u_i = i^{th} specific factor.

The unobservable random vectors $\underset{\sim}{F}$ and $\underset{\sim}{u}$ satisfy $\underset{\sim}{F}$ and $\underset{\sim}{u}$ are indpendent

$E(\underset{\sim}{u})=0$, $V(\underset{\sim}{u})=\psi=\text{doag}(\psi_1,\psi_2,\cdots,\psi_p)$

The orthogonal factor model implies a covariance structure for $\underset{\sim}{X}$ .From the model in (4) , we have

$$\left(\underset{\sim}{X}-\underset{\sim}{\mu}\right)\left(\underset{\sim}{X}-\underset{\sim}{\mu}\right)'=\left(\Lambda\underset{\sim}{F}+\underset{\sim}{u}\right)\left(\Lambda\underset{\sim}{F}+\underset{\sim}{u}\right)'$$
$$=\Lambda\underset{\sim}{F}\underset{\sim}{F}'\Lambda'+\Lambda\underset{\sim}{F}\underset{\sim}{u}'+\underset{\sim}{u}\underset{\sim}{F}'\Lambda'+\underset{\sim}{u}\underset{\sim}{u}'$$

so that

$$\boldsymbol{\Sigma}=V\left(\underset{\sim}{X}\right)=E\left(\left(\underset{\sim}{X}-\underset{\sim}{\mu}\right)\left(\underset{\sim}{X}-\underset{\sim}{\mu}\right)'\right)$$
$$=\Lambda E\left(\underset{\sim}{F}\underset{\sim}{F}'\right)\Lambda'+\Lambda E\left(\underset{\sim}{F}\underset{\sim}{u}'\right)+E\left(\underset{\sim}{u}\underset{\sim}{F}'\right)\Lambda'+E\left(\underset{\sim}{u}\underset{\sim}{u}'\right)$$
$$=\Lambda\Lambda'+\psi \qquad (\text{From(3)}) \qquad \text{---- (5)}$$

Also from the model (4),we have

$$\left(\underset{\sim}{X}-\underset{\sim}{\mu}\right)\underset{\sim}{F}'=\left(\Lambda\underset{\sim}{F}+\underset{\sim}{u}\right)\underset{\sim}{F}\quad=\Lambda\underset{\sim}{F}\underset{\sim}{F}'+\underset{\sim}{u}\underset{\sim}{F}'$$

$$\text{cov}\left(\underset{\sim}{X},\underset{\sim}{F}\right)=E\left(\left(\underset{\sim}{X}-\underset{\sim}{\mu}\right)\underset{\sim}{F}'\right)=\Lambda E\left(\underset{\sim}{F}\underset{\sim}{F}'\right)+E\left(\underset{\sim}{u}\underset{\sim}{F}'\right)=\Lambda \quad (\text{From(3)}) \quad \text{---- (6)}$$

From the model (1), we have

$$X_i=\mu_i+\sum_{j=1}^{k}\lambda_{ij}F_j+u_i\ , \qquad i=1,2,\cdots,p$$

Covariance-structure for the orthogonal factor model

$$\left.\begin{array}{l} 1.\ V\left(\underset{\sim}{X}\right)=\Lambda\Lambda'+\psi \\ \quad or\ X_i=\mu_i=\sum_{j=1}^{k}\lambda_{ij}F_j+u_i \\ \quad V\left(X_i\right)=\sum_{j=1}^{k}\lambda_{ij}^2+\psi_i\ and\ \text{cov}\left(X_i,X_l\right)=\sum_{j=1}^{k}\lambda_{ij}\lambda_{lj} \\ 2.\ \text{cov}\left(\underset{\sim}{X},\underset{\sim}{F}\right)=\Lambda\ \ or\ \ \text{cov}\left(X_i,F_j\right)=\lambda_{ij} \end{array}\right\} \qquad \text{---- (7)}$$

From the above ,thus $V(X_i)$ can be split into two parts.

First $h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2$ is called the *communality* and represents the variance of X_i which is shared with the other variables via the common factors.

In particular $\lambda_{ij}^2 = \left[\text{cov}(X_i, F_j)\right]^2$ represents the extent to which X_i depends on the j^{th} common factor . On the other hand ψ_i is called specific or unique variance and is due to the specific factor u_i it explains the variability in X_i not shared with other variables.

Thus from (7)

$$\underbrace{\sigma_{ii}}_{V(X_i)} = \underbrace{\lambda_{i1}^2 + \lambda_{i2}^2 + \cdots \lambda_{ik}^2}_{\text{communality}} + \underbrace{\psi_i}_{\text{specific variance}} \qquad \textbf{---- (8)}$$

$$\underbrace{\sigma_{ii}}_{V(X_i)} = \underbrace{h_i^2}_{\text{communality}} + \underbrace{\psi_i}_{\text{specific variance}} \qquad \textbf{---- (9)}$$

so that the i^{th} communality is the sum of squares of the loadings of the i^{th} variable on k common factors.

NOTE: The validity of the k-factor model can be expressed in terms of a simple condition on $\boldsymbol{\Sigma}$

From (5) we have

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi \qquad \textbf{---- (10)}$$

The converse also holds . If $\boldsymbol{\Sigma}$ can be decomposed into the form (10), then the k-factor model holds For $\underset{\sim}{\mathbf{X}}$. However , $\underset{\sim}{\mathbf{F}}$ and $\underset{\sim}{\mathbf{u}}$ are not uniquely determined by $\underset{\sim}{\mathbf{X}}$.

Factor analysis is invariant of scaling of variables

Suppose $\underset{\sim}{\mathbf{X}} = \underset{\sim}{\boldsymbol{\mu}} + \Lambda_x \underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{u}}$ **---- (1)**

is the factor model.

Now rescaling the variables of $\underset{\sim}{\mathbf{X}}$ is equivalent to set

$$\underset{\sim}{\mathbf{Y}} = C\underset{\sim}{\mathbf{X}}, \textit{where } \text{C=diag}(c_1, c_2, \cdots, c_p)$$

$$= \begin{pmatrix} c_1 & 0 & 0 & \cdots & 0 \\ 0 & c_1 & 0 & \cdots & 0 \\ 0 & 0 & c_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & c_1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ \vdots \\ X_p \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} c_1 X_1 \\ c_2 X_2 \\ \vdots \\ \vdots \\ c_p X_p \end{pmatrix} \qquad \textbf{- ---- (2)}$$

Premultiplying (1) with C we get

$$\underset{\sim}{\mathbf{Y}} = C\underset{\sim}{\boldsymbol{\mu}} + C\Lambda\underset{\sim}{\mathbf{F}} + C\underset{\sim}{\mathbf{u}}$$

$$\textit{and } V(\underset{\sim}{\mathbf{Y}}) = C\Lambda_x\Lambda_x'C' + C\psi_x C'$$

$$\textit{i.e., } \Sigma_y = \Lambda_y\Lambda_y' + \psi_y \qquad \text{-------(3)}$$

when $\Lambda_y = C\Lambda_x$

$$\psi_y = C\psi_x C \qquad (\because C = C')$$

$$= \text{diag}(c_1^2\psi_1, c_2^2\psi_2, \cdots, c_p^2\psi_p)$$

From (1)

$V(\underset{\sim}{\mathbf{X}}) = \Lambda_x\Lambda_x' + \psi_x$

$$\Rightarrow \Sigma_x = \Lambda_x\Lambda_x' + \psi_x \qquad \text{-------(4)}$$

But we have

$$\Sigma_y = C\Sigma_x C \qquad (\because C = C')$$

$$= C\Lambda_x\Lambda_x'C' + C\psi_x C$$

$$= \Lambda_y\Lambda_y' + \psi_y$$

Which is nothing but (3).

Thus the factor loading matrix Λ_y for the scaled random vector $\underset{\sim}{\mathbf{Y}}$ is obtained by scaling the factor loading matrix Λ_x of the original random vector $\underset{\sim}{\mathbf{X}}$. Similarly the specific variance matrix ψ_y for the scaled random vector $\underset{\sim}{\mathbf{Y}}$ is obtained by premultiplying and postmuliplying the specific variance matrix ψ_x of the original r.v. $\underset{\sim}{\mathbf{X}}$ by C . In other words,factor analysis(unlike principal component analysis) is unaffected by a rescaling of the variables.

Non-uniqueness of factor loadings(Rotated Factors)

Let T is any $k \times k$ orthogonal matrix . So that , $TT' = T'T = I$ .Then the factor model

$$\underset{\sim}{\mathbf{X}} = \underset{\sim}{\boldsymbol{\mu}} + \Lambda\underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{u}} \qquad \textbf{---- (1)}$$

Can be written as

$$\underset{\sim}{\mathbf{X}} = \underset{\sim}{\boldsymbol{\mu}} + \Lambda TT'\underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{u}}$$

$$= \underset{\sim}{\boldsymbol{\mu}} + (\Lambda T)(T'\underset{\sim}{\mathbf{F}}) + \underset{\sim}{\mathbf{u}}$$

$$= \underset{\sim}{\boldsymbol{\mu}} + \Lambda^*\underset{\sim}{\mathbf{F}}^* + \underset{\sim}{\mathbf{u}} \qquad \text{----(2)}$$

where , $\Lambda^* = \Lambda T$ and $\underset{\sim}{\mathbf{F}}^* = T'\underset{\sim}{\mathbf{F}}$

Since , $E(\underset{\sim}{\mathbf{F}}^*) = T'E(\underset{\sim}{\mathbf{F}}) = \underset{\sim}{\mathbf{0}}$ and $V(\underset{\sim}{\mathbf{F}}^*) = T'V(\underset{\sim}{\mathbf{F}})T = T'IT = I$.

It is impossible , on the basis of observations on $\underset{\sim}{\mathbf{X}}$ to distinguish the loadings Λ from those of Λ^* . That is the factor $\underset{\sim}{\mathbf{F}}$ and $\underset{\sim}{\mathbf{F}}^* = T'\underset{\sim}{\mathbf{F}}$ have the same statistical properties and even though the loadings Λ^* are in general different from the loadings Λ , they both generate the same covariance matrix. That is

$$\Sigma = \Lambda\Lambda' + \psi \qquad \text{-----(3)}$$
$$= \Lambda TT'\Lambda' + \psi$$
$$= \Lambda^{*}\Lambda^{*\prime} + \psi \qquad \text{-----(4)}$$

Thus the variance-covariance matrix $\mathbf{\Sigma}$ can be decomposed as either (3) or (4).And if Λ is the factor loadings, then $\Lambda^{*} = \Lambda T$ (for any orthogonal matrix T) , is also the factor loadings. However, the communalities given by the diagonal elements of $\Lambda\Lambda' = \Lambda^{*}\Lambda^{*\prime}$ are uneffected by the choice of T.

This determinancy in the definition of factor loadings is usually resolved by rotating (multiplying by an orthogonal matrix). The factor loadings Λ to satisfy an arbitrary constant such as $\Lambda'\psi^{1}\Lambda$ is diagonal or $\Lambda' D^{-1}\Lambda$ is diagonal , $D = diag(\sigma_{11}, \sigma_{22}, ..., \sigma_{pp})$. Where in either case the diagonal elements are written in decreasing order.. Once the loadings and specific variances are obtained , factors are identified and estimated values for the factors themselves (called factor scores) are frequently construted.

METHODS OF ESTIMATION:-

Given observations $\underset{\sim}{\mathbf{X}}_{\mathbf{1}}, \underset{\sim}{\mathbf{X}}_{2}, ..., \underset{\sim}{\mathbf{X}}_{n}$ on p generally correlated variables , factor analysis seeks to the question ,'Does the factor model

$$\underset{\sim}{\mathbf{X}} = \underset{\sim}{\boldsymbol{\mu}} + \Lambda\underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{u}} \qquad \text{----(1)}$$

With a small number of factors , adequately represent the data?

In essence , we tackle this statistical model building problem by trying to verify the covariance relationship

$$\Sigma = \Lambda\Lambda' + \psi \qquad \text{----(2)}$$

The sample covariance matrix S is an estimator of the unknown $\mathbf{\Sigma}$ .If the off-diagonal elements are small or those of the sample correlation matrix R are essentially zero,the variables are not related and factor analysis will not prove useful.In these circumstances ,the specific factors play the dominant role ,where as the major aim of the factor analysis is to determine a few important common factors.

If $\mathbf{\Sigma}$ appears to deviate significantly from diagonal matrix then a factor model can be entertained and the initial problem is one of the estimating the factor loadings $\lambda_{ij}'s$ and specific variances $\psi_{i}'s$ .We shall consider two of the most popular methods of parameter estimation.

1.Principal factor method (Analysis).

2. Maximum likelihood method (factor analysis)

The solution from either method can be rotated in order to simplify the interpretation of factors

Principal factor analysis:

We have the factor model (k-factor)

$$\underset{\sim}{\mathbf{X}}_{(p\times 1)} = \underset{\sim}{\boldsymbol{\mu}}_{(p\times 1)} + \Lambda_{(p\times k)}\underset{\sim}{\mathbf{F}}_{(k\times 1)} + \underset{\sim}{\mathbf{u}}_{(p\times 1)} \qquad \text{-----(1)}$$

Where $\underset{\sim}{\mathbf{X}}$ = p-component random vector

$\underset{\sim}{\boldsymbol{\mu}}$ = mean of $\underset{\sim}{\mathbf{X}}$

$\boldsymbol{\Lambda}$ = matrix of factor loadings

$\underset{\sim}{\mathbf{F}}$ = vector of common factors

$\underset{\sim}{\mathbf{u}}$ = p-component random vector

with covariance matrix of $\underset{\sim}{\mathbf{X}}$

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\psi} \qquad \text{-----(2)}$$

Where $\boldsymbol{\psi} = V(\underset{\sim}{\mathbf{u}}) = diag\left(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \cdots, \boldsymbol{\psi}_k\right)$.

In practical situation ,since $\boldsymbol{\Sigma}$ is not known, $\boldsymbol{\Sigma}$ is replaced by its estimate the sample covariance matrix $\mathbf{S}$ which is obtained from the observations $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ .Since,factor analysis is invariant of the scaling of the variables the correlaton matrix $\mathbf{R}$,computed from the observations $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ on p-variable random vector $\underset{\sim}{\mathbf{X}}$,may also be used in place of $\mathbf{S}$.

Let us suppose the data is summerised by the correlation matrix $\mathbf{R}$ so that an estimate of $\boldsymbol{\Lambda}$ and $\boldsymbol{\psi}$ is rought for the standardised variables.

Now our problem is to obtain the estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\psi}$ from equation (2) ,replacing the unknown $\boldsymbol{\Sigma}$ with known $\mathbf{R}$ (when the variables standardised $\boldsymbol{\Sigma}$ is equivalent to the population correlation matrix $\boldsymbol{\rho}$).Then we have

$$\mathbf{R} = \hat{\Lambda}\hat{\Lambda}' + \hat{\boldsymbol{\psi}} \qquad \text{-----(3)}$$

Comparing the diagonal elements on both sides ,we get

$$1 = \hat{h}_i^2 + \hat{\psi}_i \quad \text{for i=1,2,---,p}$$

$$where \quad \hat{h}_i^2 = \sum_{j=1}^{k} \hat{\boldsymbol{\lambda}}_{ij}^2$$

Is the preliminaryestimate of the i^{th} communality h_i^2 and may be obtained either of the following two ways:

1) The square of the multiple correlation coefficient of the i^{th} variable X_i on the remaining p-1 variables.

2) The largest absolute correlation coefficient between X_i and one of the remaining p-1 variables.

i.e., $\max_{j \neq i} |r_{ij}|$

Note that the estimated communality h_i^2 is higher when X_i is highly correlated with the other as we would expect. Now $\hat{\boldsymbol{\psi}}=\text{diag}(\hat{\boldsymbol{\psi}}_i) = diag\left(1-\hat{h}_i^2\right)$ has to be subtracted from $\mathbf{R}$ to obtained the matrix

$$\mathbf{R}-\hat{\boldsymbol{\psi}} = \begin{bmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ r_{12} & \hat{h}_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{1p} & r_{2p} & \cdots & \hat{h}_p^2 \end{bmatrix} \qquad \text{---- (4)}$$

Which is called the reduced correlation matrix because the 1's on the diagonal have been replaced by the estimated *communalities* $\hat{h}_i^2$.

Suppose $a_1 \geq a_2 \geq \cdots \geq a_p$ are eigen values of $\mathbf{R}-\hat{\boldsymbol{\psi}}$ and $\underset{\sim}{\boldsymbol{\omega}}_1, \underset{\sim}{\boldsymbol{\omega}}_2, \cdots, \underset{\sim}{\boldsymbol{\omega}}_p$ are the corresponding eigen vectors, then we may decompose $\mathbf{R}-\hat{\boldsymbol{\psi}}$ as

$$\mathbf{R}-\hat{\boldsymbol{\psi}} = \sum_{i=1}^{p} a_i \underset{\sim}{\boldsymbol{\omega}}_i \underset{\sim}{\boldsymbol{\omega}}_i' \qquad \text{---- (5)}$$

Suppose the first K eigen values $a_1, a_2, \cdots, a_k$ are positive then

$$\mathbf{R}-\hat{\boldsymbol{\psi}} = \sum_{i=1}^{p} a_i \underset{\sim}{\boldsymbol{\omega}}_i \underset{\sim}{\boldsymbol{\omega}}_i' = \hat{\Lambda}\hat{\Lambda}' \qquad \text{---- (6)}$$

Where, $\hat{\Lambda} = \left[\sqrt{a_1}\,\underset{\sim}{\boldsymbol{\omega}}_1 \quad \sqrt{a_2}\,\underset{\sim}{\boldsymbol{\omega}}_2 \quad \cdots \quad \sqrt{a_k}\,\underset{\sim}{\boldsymbol{\omega}}_k\right]_{p\times k}$

$$= \boldsymbol{\Omega}\mathbf{A}^{\frac{1}{2}} \qquad \text{---- (7)}$$

$\Omega = \left(\underset{\sim}{\boldsymbol{\omega}}_1 \quad \underset{\sim}{\boldsymbol{\omega}}_2 \quad \cdots \quad \underset{\sim}{\boldsymbol{\omega}}_k\right)$ and $\mathbf{A} = diag\left(a_1, a_2, \cdots, a_k\right)$ is the estimate of the factor loading matrix $\hat{\boldsymbol{\Lambda}}$. Since $\boldsymbol{\Omega}$ is orthogonal matrix, we may see that

$$\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Lambda}} = \mathbf{A}^{1/2}\boldsymbol{\Omega}'\boldsymbol{\Omega}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{I}\mathbf{A}^{1/2} = \mathbf{A} \qquad \text{---- (8)}$$

Finally, the revised estimates of the specific variances are given in terms of $\hat{\mathbf{\Lambda}}$ by

$$\hat{\boldsymbol{\psi}}_{\mathbf{i}} = 1 - \sum_{j=1}^{k} \hat{\boldsymbol{\lambda}}_{ij}^2, i = 1,2,\cdots,p \qquad \textbf{---- (9)}$$

Where $\hat{\boldsymbol{\lambda}}_{ij}$ is the $(i,j)^{th}$ element of the estimated factor loading s matrix $\hat{\mathbf{\Lambda}}$ given by **(7)** .Then the principal factor solution is permissible is all the $\hat{\boldsymbol{\psi}}_{\mathbf{i}}$ are non-negetive.

Thus for the k factor model (1) the principal factor estimates of the factors loading matrix $\mathbf{\Lambda}$ is given by (7) and the estimates of communalities h_i^2 are given by the diagonal elements of $\hat{\Lambda}\hat{\Lambda}'$.

i.e. $$h_i^2 = \sum_{j=1}^{k} \hat{\boldsymbol{\lambda}}_{ij}^2 \qquad \textbf{---- (10)}$$

The estimates of the specific variables $\hat{\boldsymbol{\psi}}_{\mathbf{i}}'$s are given by (9)

NOTE:

- The principal factor analysis can be performed iteratively with the communality estimates given by (10)
- becoming the initial estimates for the next stage.
- If we are given the sample covariance matrix **S** ,it may be converted into **R** and then above analysis can be performed.

Example:

Consider the open/closed book data of the following table with correlation matrix.

$$\begin{bmatrix} 1 & 0.553 & 0.547 & 0.410 & 0.389 \\ & 1 & 0.610 & 0.485 & 0.437 \\ & & 1 & 0.711 & 0.665 \\ & & & 1 & 0.607 \\ & & & & 1 \end{bmatrix}$$

If k>2 then **S**<0 and the factor model is not well defined . The principal factor solutions for k=1 and k=2, where we estimate the i[th] communality $\hat{h}_i^2$ by $\max_j |r_{ij}|$,are given in the table . The eigen values of the reduced correlation matrix are 2.84,0.38, 0.08,0.02 and -0.05, suggesting that the two-factor solution fits the data well.

In the above table principal factor solutions for the open/closed book data with k=1 and k=2 factors.

variable	$\overbrace{\hat{h}_i^2 \quad \lambda(1)}^{k=1}$		$\overbrace{\hat{h}_i^2 \quad \lambda(1) \quad \lambda(2)}^{k=2}$		
1	0.417	0.646	0.543	0.646	0.354
2	0.506	0.711	0.597	0.711	0.303
3	0.746	0.864	0.749	0.864	-0.051
4	0.618	0.786	0.680	0.786	-0.249
5	0.551	0.742	0.627	0.742	-0.276

The first factor represents overall performance and for k=2, the second factor , which is much less importent

$\left(a_2 = 0.38 << 2.84 = a_1\right)$, represents a contrast acrossthe range $h_i^2 << 1$ for all i, and therefore a fair propertion of the variance of each variable is left unexplained

by the comman factor.

PRINCIPAL COMPONENT METHOD (PRINCIPAL COMPONENT SOLUTIO OF THE FACTOR MODEL)

Suppose $\underset{\sim}{\mathbf{x}}_1, \ldots, \underset{\sim}{\mathbf{x}}_n$ are observations on **p** generally correlated variables and the data is summarised either into the sample correlation matrix **R**.

Let the orthogonal factor model with k comman factors

$$\underset{\sim}{\mathbf{X}}_{(p\times 1)} = \underset{\sim}{\boldsymbol{\mu}}_{(p\times 1)} + \Lambda_{(p\times k)}\underset{\sim}{\mathbf{F}}_{(k\times 1)} + \underset{\sim}{\mathbf{u}}_{(p\times 1)} \qquad \text{-----(1)}$$

Where $\underset{\sim}{\mathbf{X}}$ = p-component random vector

$\underset{\sim}{\boldsymbol{\mu}}$ = mean of $\underset{\sim}{\mathbf{X}}$

$\boldsymbol{\Lambda}$ = matrix of factor loadings

$\underset{\sim}{\mathbf{F}}$ = vector of common factors

$\underset{\sim}{\mathbf{u}}$ = vector of random distrubances

with $\quad V(\underset{\sim}{\mathbf{X}}) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\psi}$ **-----(1.a)**

$$\boldsymbol{\psi} = V(\underset{\sim}{\mathbf{u}})$$

Now the principal component method is to obtain the estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\psi}$ using the sample covariance matrix **S** or sample correlated matrix **R** .

Suppose $a_1 \geq a_2 \geq \cdots \geq a_p$ are the latent roots of **S (**or **R**) and let us consider the first 'k' roots

i.e. $a_1, a_2, \cdots, a_p$

Let $\underset{\sim}{\boldsymbol{\omega}}_1, \underset{\sim}{\boldsymbol{\omega}}_2, \cdots, \underset{\sim}{\boldsymbol{\omega}}_k$ be the corresponding latent vectors. Then the estimated matrix of factor loadings is given by

$$\hat{\boldsymbol{\Lambda}} = \left[\sqrt{a_1}\,\underset{\sim}{\boldsymbol{\omega}}_1 \quad \sqrt{a_2}\,\underset{\sim}{\boldsymbol{\omega}}_2 \quad \cdots \quad \sqrt{a_k}\,\underset{\sim}{\boldsymbol{\omega}}_k\right]_{p\times k} \qquad \text{-----(2)}$$

and the estimated specific variances are provided by the diagonal elements of the matrix

$$\mathbf{S} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'\left(\mathbf{R} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'\right) \qquad \text{-----(3)}$$

so that $\hat{\boldsymbol{\psi}}=\text{diag}\left(\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_2, \cdots, \hat{\boldsymbol{\psi}}_k\right)$ with $\hat{\boldsymbol{\psi}}_i = s_{ii} - \sum_{j=1}^{k} \hat{\boldsymbol{\lambda}}_{ij}^2$ $\left(\hat{\boldsymbol{\psi}}_i = r_{ii} - \sum_{j=1}^{k} \hat{\boldsymbol{\lambda}}_{ij}^2\right)$ estimates of communalities are given by the diagonal elements of $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$

i.e.

$$h_i^2 = \sum_{j=1}^{k} \hat{\lambda}_{ij}^2$$

NOTE:

- Consider the residual matrix $\mathbf{S} - \left(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\psi}}\right)$ ersulting from the approximation of **S** by the principal component solution. The diagonal elements are zero and if the other elements are also small,we may subjectively take the 'k' factor model to be appropriate.
- The contribution to the total sample variance = Tr(**S**) from the j^{th} commaon factor is given by

$$\sum_{i=1}^{p} \hat{\boldsymbol{\lambda}}_{ij}^2 = \left(\sqrt{a_j}\ \underset{\sim}{\boldsymbol{\omega}}_j\right)'\left(\sqrt{a_j}\ \underset{\sim}{\boldsymbol{\omega}}_j\right)$$

$$= a_j \qquad \left(\because\ \underset{\sim}{\boldsymbol{\omega}}_j'\underset{\sim}{\boldsymbol{\omega}}_j = 1\right)$$

$$= j^{th} \text{ latent root of } \mathbf{S} \text{ (Where is } a_j \text{ the } j^{th} \text{ latent root of } \mathbf{S})$$

Thus , proportion of total sample variance due to j^{th} factor

$$= \begin{cases} \dfrac{a_j}{\text{Tr(S)}} & \text{for factor analysis of 'S'} \\ \dfrac{a_j}{\text{p}} & \text{for factor analysis of 'R'} \end{cases}$$

- (For worked out examples see page nos: 388-391 of Applied Multivariate Analysis by Johnson&Wichern)

MAXIMUM LIKELYHOOD FACTOR ANLYSIS:

Suppose $\underset{\sim}{\mathbf{x}}_1,\ldots,\underset{\sim}{\mathbf{x}}_n$ are 'n' observations drawn on $\underset{\sim}{\mathbf{X}}$ which follows population $Np\left(\underset{\sim}{\boldsymbol{\mu}},\boldsymbol{\Sigma}\right)$ and $\underset{\sim}{\mathbf{X}}$ is having the following k factor model

$$\underset{\sim}{\mathbf{X}}_{(p\times 1)} = \underset{\sim}{\boldsymbol{\mu}}_{(p\times 1)} + \Lambda_{(p\times k)}\underset{\sim}{\mathbf{F}}_{(k\times 1)} + \underset{\sim}{\mathbf{u}}_{(p\times 1)}$$

$\underset{\sim}{\boldsymbol{\mu}}$ = mean of $\underset{\sim}{\mathbf{X}}$

$\boldsymbol{\Lambda}$ = matrix of factor loadings

$\underset{\sim}{\mathbf{F}}$ = vector of common factors

$\underset{\sim}{\mathbf{u}}$ = vector of random distrubances

with the asumptions

$E(\underset{\sim}{F})=0=E(\underset{\sim}{u})$

$V(\underset{\sim}{F})=I$, $V(\underset{\sim}{u})=\psi=\text{doag}(\psi_1,\psi_2,\cdots,\psi_p)$

$\text{cov}(\underset{\sim}{F},\underset{\sim}{u})=0$

These assumptions implicitly imposes the restriction on $\boldsymbol{\Sigma}$ as follows

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi \qquad \text{-----(2)}$$

Since $\underset{\sim}{\mathbf{X}} \sim Np\left(\underset{\sim}{\boldsymbol{\mu}},\boldsymbol{\Sigma}\right)$, its log -likelyhood is given by

$$\log L = \frac{-n}{2}\log\left|2\pi\boldsymbol{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{n}\left(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}\right)'\boldsymbol{\Sigma}^{-1}\left(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}\right)$$

if we with its MLE $\bar{\mathbf{X}}$, then $\log L$ becomes

$$l = \log L = \frac{-n}{2}\log\left|2\pi\boldsymbol{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{n}\left(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}\right)'\boldsymbol{\Sigma}^{-1}\left(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}\right)$$

$$= -\left(\frac{n}{2}\log\left|2\pi\boldsymbol{\Sigma}\right| + \frac{\text{n}}{2}\ \text{Tr}(\boldsymbol{\Sigma}^{-1}\text{S}_\text{n})\right) \qquad \text{-----(3)}$$

Where

$$S_n = \frac{1}{n}\sum_{i=1}^{n}\left(\underset{\sim}{x}_i - \underset{\sim}{\mu}\right)'\left(\underset{\sim}{x}_i - \underset{\sim}{\mu}\right)$$

Σ is as given by **(2)**

Maximizing (3) is equivalent to minimizing the following function w.r.t. Λ and ψ

$$F(\Lambda,\bar{\psi}) = \log|\Sigma| + \mathrm{Tr}\left(\Sigma^{-1}S_n\right) - \log|S_n| - p \qquad \text{----- (4)}$$

$$\left(\because |S_n| \text{ and p are constants}\right)$$

Since from (2) , Λ is not uniquely determined . We have minimize (4) subject to the following uniqueness condition

$$\Lambda'\psi^{-1}\Lambda = \Delta, \text{a diagonal matrix} \qquad \text{-----(5)}$$

The MLEs $\hat{\Lambda}$ and $\hat{\psi}$ obtained by minimizing (4) subject to (5) satisfy

$$\left(\hat{\psi}^{\frac{-1}{2}}S_n\hat{\psi}^{\frac{-1}{2}}\right)\left(\hat{\psi}^{\frac{-1}{2}}\hat{\Lambda}\right) = \left(\hat{\psi}^{\frac{-1}{2}}\hat{\Lambda}\right)\left(1+\hat{\Delta}\right) \qquad \text{-----(6)}$$

so that the j^{th} column of $\hat{\psi}^{\frac{-1}{2}}\hat{\Lambda}$ is the (non-normalised) eigen vector of $\hat{\psi}^{\frac{-1}{2}}S_n\hat{\psi}^{\frac{-1}{2}}$ corresponding to eigen value $1+\hat{\Delta}_i$

where $\hat{\Delta}_1 \geq \hat{\Delta}_2 \geq \cdots \geq \hat{\Delta}_k$

clearly,for the above, the MLE of $\hat{\Lambda}$ can be obtained only for a given $\hat{\psi}$,whose initial value can be taken as

$$\hat{\psi}^{(0)} = diag\left(\hat{\psi}_1^{(0)}, \hat{\psi}_2^{(0)}, \cdots, \hat{\psi}_k^{(0)}\right)$$

where $\hat{\psi}_i = \left(1 - \frac{1}{2}\frac{k}{p}\right)\left(\frac{1}{s^{ii}}\right)$

where s^{ii} is the i^{th} diagonal element of $\mathbf{S} = \frac{n\,\mathbf{S}_n}{(n-1)}$

The next modified value of $\hat{\psi}$ is given by

$$\hat{\psi}^{(1)} = diag\left(\hat{\psi}_1^{(1)}, \hat{\psi}_2^{(1)}, \cdots, \hat{\psi}_k^{(1)}\right)$$

Where $\hat{\psi}^{(1)}$ is the i^{th} diagonal element of the computed matrix $\mathbf{S}_n - \hat{\Lambda}\hat{\Lambda}'$

Using this $\hat{\psi}^{(1)}$,we can obtained the revised value of $\hat{\Lambda}$ using (6).

This procedure is to be continued until the latest estimates $\hat{\Lambda}$ and $\hat{\psi}$ satisfy the relation (5).

NOTE:

Ordinarily the observations are standardised and a sample correlation matrix is factor analysed.Of the data is

summerised into a sample correlation matrix $\mathbf{R}$,then the above method of maximum likelihood factor analysis may be

carried out replacing $\mathbf{S}_n$ or $\mathbf{S}$ by $\mathbf{R}$ to get the some estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\psi}$ .This is due to the fact that the MLEs are scale invariant.

An worked out example is given in page no:394 of Applied Multivariate Statistical analysis by Richard A.Jhon and Wichern.

GOODNES OF FIT TEST:

One of the main advantages of the maximum likelihood technique is that it provides a test of the hypothesis

H_0 : k comman factors are sufficient to against describe the data .

H_1 : $\boldsymbol{\Sigma}$ has no constraints.

The likelihood ratio statistic λ is given by

$$-2\log\lambda = np(\hat{a} - \log\hat{g} - 1) \qquad \text{-----(1)}$$

where $\hat{a}$ and $\hat{g}$ are the arithmatic and geomatric means of the eigen values of $\boldsymbol{\Sigma}^{-1}\mathrm{S}_n$.

But we have

$$\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathrm{S}\right) - \log\left|\hat{\boldsymbol{\Sigma}}^{-1}\mathrm{S}\right| - \mathrm{p} = \mathrm{p}\left(\hat{\mathrm{a}} - \log\hat{\mathrm{g}} - 1\right) \qquad \text{-----(2)}$$

When $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\psi}}$

comparing (1) & (2), we have

$$-2\log\lambda = n\left(\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathrm{S}\right) - \log\left|\hat{\boldsymbol{\Sigma}}^{-1}\mathrm{S}\right| - \mathrm{p}\right)$$

$$\text{when} \quad \hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\psi}}$$

Now under H_0 the statistic $-2\log\lambda$ has an asymptotic χ^2_m distribution ,

where

$$\mathrm{m} = \frac{1}{2}(p-k)^2 - \frac{1}{2}(p-k)$$

According to Bartlet, an improved test statistic is given by (3) replacing n by

$$\mathrm{n}' = \left(\mathrm{n} - 1 - \left(2p + 4k + 5\right)/6\right)$$

NOTE: See for example 9.7 page no.399 of AMSA by Johnson.

FACTOR ROTATION:

We have

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\psi} \\ &= \boldsymbol{\Lambda}\mathrm{T}\mathrm{T}'\boldsymbol{\Lambda}' + \boldsymbol{\psi} \\ &= (\boldsymbol{\Lambda}\mathrm{T})(\boldsymbol{\Lambda}\mathrm{T})' + \boldsymbol{\psi} \\ &= \boldsymbol{\Lambda}^{*}(\boldsymbol{\Lambda}^{*})' + \boldsymbol{\psi}\end{aligned}$$

where $\boldsymbol{\Lambda}^{*} = \boldsymbol{\Lambda}\mathrm{T}$

T is an orthogonal matrix

Thus if $\boldsymbol{\Lambda}$ is a factor loadings matrix which reproduce $\boldsymbol{\Sigma}$,then any other factor loadings matrix $\boldsymbol{\Lambda}^{*}$ obtained from $\boldsymbol{\Lambda}$ by an orthogonal transformation (T) have the same ability to reproduce the covariance matrix (or correlation matrix).From matrix algebra,we know that an orthogonal transformation corresponds to a rigid rotation of the coordinate axes.For this reasonan orthogonal transformation of the factor loadings and the implied orthogonal transformation of the factor is called "factor rotation".

Let $\hat{\boldsymbol{\Lambda}}$ be the $p \times k$ matrix of estimated factor loadings obtained by any method ,then

$$\hat{\boldsymbol{\Lambda}}^{*} = \hat{\boldsymbol{\Lambda}}\mathrm{T} \quad \text{where } \mathrm{T}\mathrm{T}'=\mathrm{T}'\mathrm{T}=\mathrm{I} \qquad \text{-----(1)}$$

is a $p \times k$ matrix of rotated loadings. Moreover ,the estimated covariance (or correlation) matrix remains unchange ,

since $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\psi}} = \hat{\boldsymbol{\Lambda}}^{*}(\hat{\boldsymbol{\Lambda}}^{*})' + \hat{\boldsymbol{\psi}}$ **-----(2)**

since ,the original loadings may not be readily interpretable , it is usual practice to rotate them until a " sample structure" is achieved.Ideally we should like to see a patern of loadings such that each variable loadings highly on a single factor and has

small-to-moderate loadings on the remaining factors. Of course, it is not always possible to get this simple structure.A conveniant analytical choice of rotation is given by the "varimax method" described below:The verimax method of orthogonal rotation was provided by **kaiser**(1958).Its rationale is to provide axes with a few large loadings and as many near zero loadings as possible.This is accomplished by an iterative maximization of a quadratic function of the loadings.

Devote the matrix of rotated loadings as

$$\hat{\Delta} = \hat{\mathbf{\Lambda}}\mathrm{T}$$

Now the $(i,j)^{th}$ element of $\mathbf{\Delta}$ viz; δ_{ij} represants the loadings of the i^{th} variable on the j^{th} factor .

The function ϕ that the variance criterionmaximizes is the sum of the variances of the squared loadings within each

column of the loadings is normalised by its communality,that is

$$\phi = \sum_{i=1}^{k}\sum_{j=1}^{p}\left(d_{ij}^2 - \bar{d}_i\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} d_{ij}^4 - p\sum_{i=1}^{k}\bar{d}_i^{\,2}$$

Where $d_{ij} = \dfrac{\delta_{ij}}{h_i}$ *and* $\bar{d} = \dfrac{1}{p}\sum_{j=1}^{p} d_{ij}^2$

h_i^2 is the i^{th} communality is the i^{th} diagonal element of $\hat{\Delta}\,\hat{\Delta}'$

The varimax criterian ϕ is a function of T , and the iterative algorithm proposed by Kariser finds the orthogonal matrix

$\mathbf{G}$ which maximizes ϕ.

In the case where k=2,the calculations simplify.For then T is given by

$$\mathrm{T}=\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

and represents a rotation of the coordinate axis clockwise by an angle θ . The value of θ can be determined by

the relation $\mathrm{T}'\mathrm{T}=\mathrm{I}$

In the case where k>2 ,an iterative solution for the rotation is used.

See example 9.8,9.9,9.10,9.11 in the pages 401-408 of AMVA by Richard Johanson & Wichern.

DISCRIMINATION AND CLASSIFICATION

INTRODUCTION:

Discriminant analysis and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a seperatory procedure, it is often employed on a onetime basis in order to investigate observed differences when causal relationships are not well understood. Classification procedure are less exploratory in the sense that they lead to well defined rules, which can be used for assigning new objects. Classification ordinarily requires more problem structure than discrimination. Thus, the immediate goals of discrimination and classification, respectively, are as follows:

Goal 1: To describe either graphically (in three or fewer dimensions) or algebraically, the differential features of objects (observations) from several known collections(populations). We try to find "discriminants" whose numerical values are such that the collection are separated as much as possible.

Goal 2: To sort objects(observations) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign a new object to the labeled classes.

We shall follows convention and use the term discrimination to refer to 'Goal 1'. This terminology was introduced by R.A.Fisher in the first modern treatment of separatory problems. A more descriptive term for this goal, however, is separation. We shall refer to the second goal as classification, or allocation. A function that separates may some times serve as an allocator, and, conversely, an allocator rule may suggest a discriminatory procedure. Thus, in practice Goal 1& Goal2 frequently overlap and distinction between separation and allocation is not clear.

The problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of several categories on the basis of these measurements. The investigator cannot identify the individual with a category directly but must use these measurements. In many cases it can be assumed that these are a finite number of categories or populations from which the individual may have come and each population is characterised by a probability distribution of the measurements. Thus, an individual is considered as a random observation from this population. The question is : Given an individual with certain measurements, from which population did it arise?

In some, instances, the categories are specified before hand in the sense that the probability distributions of the measurements are completely known. In other cases, the form of each distribution may be known, but the parameters of the distribution must be estimated from a sample from that population .In some other cases, the form of the distribution of the populations may not be known.

Let us give an example of a problem of discrimination and classification. Prospective students applying for admission into college are given a battery of tests; the vector of scores is a set of measurements $\underset{\sim}{\mathbf{x}}$. The prospective students may be a member of one population consisting of these students who will successfully complete college training or, rather, have potentialities for successfully completing training, or he/she may be member of the other population, those who will not complete the course successfully. The problem is to classify a student applying for admission on the basis of these scores on the entrance examination. Before that we have to describe or explore the differential scores between the two categories of the students from the past information. Also, we have to prepare a discriminant function that separates the two categories of students clearly as much as possible. This problem is called discrimination.

SEPERATION AND CLASSIFICATION FOR TWO POPULATION:

To fix ideas, we list below situations where one may be interested in

(1). Separating or discriminating two classes of objects.

Or (2). Assigning a new object to one of the two classes .

Or both (1)&(2).

It is convenient to label the classes $\boldsymbol{\pi_1}$ & $\boldsymbol{\pi_2}$. The objects are ordinarily separated or classified on the basis of measurements on, for instance, P associated random variables. $\underset{\sim}{\mathbf{X}}' = (X_1, X_2, ..., X_p)$. The observed values of $\underset{\sim}{\mathbf{X}}$ differ to some extend from one class to the other (of the values of $\underset{\sim}{\mathbf{X}}$ were not very different for objects in $\boldsymbol{\pi_1}$ & $\boldsymbol{\pi_2}$,there would be no problem; i.e., the would be indistinguishable and new objects could be assigned to either class indiscriminately). We can think of the totality of values from the first class as being the population of $\underset{\sim}{\mathbf{x}}$ values for $\boldsymbol{\pi_1}$ and those from the second class as the population of $\underset{\sim}{\mathbf{x}}$ values for $\boldsymbol{\pi_2}$. These two populations can then be described by probability density functions

$f_1(\underset{\sim}{\mathbf{x}})$ & $f_2(\underset{\sim}{\mathbf{x}})$, and consequently, we can talk of assigning observations to populations (or objects to classes).

The following are some more examples:

(1). Separation of two species of chickweed based on the measurements sepal and petal lengths, petal left depth, bract length, scarious tip length and pollen diameter.

(2). Discrimination of successful and unsuccessful college students based on the entrance examination scores, high school grade point average and number of high school activities.

(3). Classification of purchasers of a new product and laggards (those slow to purchase) based on particulars of education, income, family size and amount of previous brand switching.

(4). Discriminating male-skulls and female-skulls based on the anthropological measurements like circumference and volume on ancient skulls.

(5). Separating good and poor credit risks based on the particulars of income, age, member of credit cards and family size.

From the above examples, it is clear that allocation or classification rules are usually developed from learning samples. Measured characteristics of randomly selected objects known to come from each of the two populations are examined for differences. Essentially, the set of possible sample outcomes is divided into two regions R_1 & R_2, such that is a new observation falls in R_1, it is allocated to population $\boldsymbol{\pi_1}$ and is it falls in R_2, we allocate it to population $\boldsymbol{\pi_2}$. Thus one set of observed values favours $\boldsymbol{\pi_1}$, the other set of values favours $\boldsymbol{\pi_2}$. Here, it may be noted that classification rules cannot usually provide and error-free method of assignment. This is because there may not be a clear distinction between the measured characteristics of the populations ; i.e. the groups may overlap. It is then possible, for example, to incorrectly classify a $\boldsymbol{\pi_2}$ object as belonging to $\boldsymbol{\pi_1}$ or a $\boldsymbol{\pi_1}$ object as belonging to $\boldsymbol{\pi_2}$.

A good classification procedure should result in a few misclassifications. In other words, the chances or probabilities of misclassification should be small. As we shall see, there are additional features that an "optimal" classification rule should be possessed.

STANDARDS OF GOOD CLASSIFICATION:

In constructing a procedure of classification, it is desired to minimize the probability of misclassification or more specifically it is desired to minimize on the average the bad effects of misclassification.

Suppose an individual is an observation from either population $\boldsymbol{\pi_1}$ or population $\boldsymbol{\pi_2}$. The classification of an observation depends on the vector of measurements

$$\underset{\sim}{\mathbf{x}} = \left(x_1, x_2, \ldots . x_p\right)'_{p\times 1}$$

on that individual. We set up a rule that if an individual is characterized by certain sets of values of $x_1, x_2, \ldots, x_p$ it will be classified as from $\boldsymbol{\pi_1}$; if it has other values it is classified as from $\boldsymbol{\pi_2}$.

We can think of an observation $\underset{\sim}{\mathbf{x}}$ as a point in a P-dimensional space. We divide this space into two regions $R_1 \& R_2$ if the observation falls in R_1, we classify it as coming from $\boldsymbol{\pi_1}$ and if it falls in R_2 we classify it as coming from $\boldsymbol{\pi_2}$.

Usually, the statistician can make two kinds of errors in classification. If the individual is actually from $\boldsymbol{\pi_1}$ and is misclassified into $\boldsymbol{\pi_2}$; or if it is actually from $\boldsymbol{\pi_2}$ and is misclassified into $\boldsymbol{\pi_1}$. We need to know the relative undesirability of these two kinds of misclassification.

Let $f_1(\underset{\sim}{\mathbf{x}}) \& f_2(\underset{\sim}{\mathbf{x}})$ be the p.d.f.'s associated with the $p\times 1$ random vector $\underset{\sim}{\mathbf{X}}$ for populations $\boldsymbol{\pi_1} \& \boldsymbol{\pi_2}$ respectively. An object, with associated measurements $\underset{\sim}{\mathbf{x}}$, must be assigned to either $\boldsymbol{\pi_1}$**(or)**$\boldsymbol{\pi_2}$. Let $\boldsymbol{\Omega}$ be the sample space that is the collection of all possible observations $\underset{\sim}{\mathbf{x}}$. Let R_1 be that set of $\underset{\sim}{\mathbf{x}}$ values for which we classify objects as $\boldsymbol{\pi_1}$ and $R_2 = \Omega - R_1$ be the remaining $\underset{\sim}{\mathbf{x}}$ values for which we classify objects as $\boldsymbol{\pi_2}$. Since every object must be assigned to one and only one of the two populations, the sets $R_1 \& R_2$ be mutually exclusive and exhaustive.

EXPECTED (OR AVERAGE) COST OF MISCLASSIFICATION (ECM):

In order to obtain ECM we consider the following conditional probabilities:

P (correctly classifying an observation (object) that actually is drawn from $\boldsymbol{\pi_1}$)

$$= P(\underset{\sim}{\mathbf{X}} \in R_1 / \boldsymbol{\pi_1}) = \int_{R_1} f_1(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} = \text{P(1/1) (say)} \quad \rightarrow (1)$$

P(correctly classifying an observation that actually is drawn from $\boldsymbol{\pi_2}$)

$$= P(\underset{\sim}{\mathbf{X}} \in R_2 / \boldsymbol{\pi_2}) = \int_{R_2 = \Omega - R_1} f_2(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} = \text{P(2/2)} \quad \text{(say)} \quad \rightarrow (2)$$

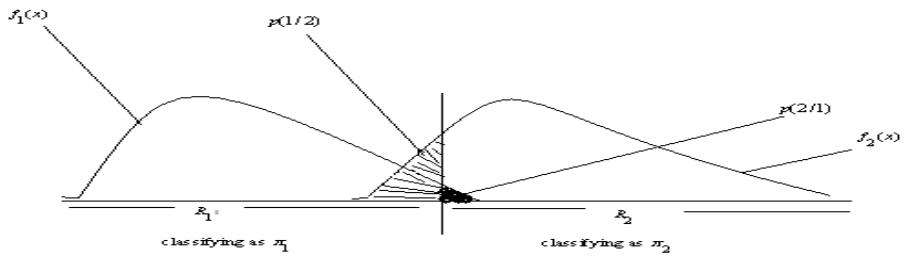
P(misclassifying an observation that is drawn from $\boldsymbol{\pi_1}$)

$$= P(\underset{\sim}{\mathbf{X}} \in R_2 / \boldsymbol{\pi}_1) = \int_{R_2} f_1(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} \quad = \text{P(2/1)} \quad \text{(say)} \quad \rightarrow (3)$$

P(misclassifying an observation that is drawn from $\boldsymbol{\pi_2}$)

$$= P(\underset{\sim}{\mathbf{X}} \in R_1 / \boldsymbol{\pi_2}) = \int_{R_1} f_2(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} \quad = \text{P(1/2)} \quad \text{(say)} \quad \rightarrow (4)$$

Misclassification probabilities when p=1:



Let

P_1 = prior probability of $\boldsymbol{\pi_1}$

= P(drawing an observation from $\boldsymbol{\pi_1}$) = P($\boldsymbol{\pi_1}$) → (5)

and P_2 = prior probability of $\boldsymbol{\pi_2}$

= P(drawing an observation from $\boldsymbol{\pi_2}$) = P($\boldsymbol{\pi_2}$) → (6)

Now the overall probabilities of correctly or incorrectly classifying objects can be derived as the product of the prior and conditional classification probabilities. Thus we get

P(correctly classified as $\boldsymbol{\pi_1}$)=P(observations comes from $\boldsymbol{\pi_1}$ and is correctly Classified as $\boldsymbol{\pi_1}$)

$= P(\underset{\sim}{\mathbf{X}} \in R_1 / \boldsymbol{\pi}_1).P(\boldsymbol{\pi}_1)$ = P(1/1). P_1 (from (1)&(5)) → (7)

similarly

P(correctly classified as $\boldsymbol{\pi}_2$)=P(2/2). P_2 (from (2)&(6)) → (8)

P(misclassified as $\boldsymbol{\pi}_1$) = P(observations comes from $\boldsymbol{\pi}_2$ and is misclassified as $\boldsymbol{\pi}_1$)

$= P(\underset{\sim}{\mathbf{X}} \in R_1 / \boldsymbol{\pi}_2).P(\boldsymbol{\pi}_2)$ = P(1/2). P_2 (from (4)&(6)) → (9)

P(misclassified as $\boldsymbol{\pi}_2$) = P(observation comes from $\boldsymbol{\pi}_1$ and is misclassified as $\boldsymbol{\pi}_2$)

$= P(\underset{\sim}{\mathbf{X}} \in R_2 / \boldsymbol{\pi}_1).P(\boldsymbol{\pi}_1)$ = P(2/1). P_1 (from (3)&(5)) → (10)

A good classification rule must take into account the misclassification costs. Although the statistician may not know these costs in each case, he will often have at least a rough idea of them. The costs of misclassification can be defined by a cost matrix C:

True population| Classified as

	π_1	π_2
π_1	0	C(2/1)
π_2	C(1/2)	0

→(11)

The costs are

(1). Zero for correct classification .

(2). C(1/2) is cost involved when an observation drawn from $\boldsymbol{\pi}_2$ is incorrectly classified into $\boldsymbol{\pi}_1$.

(3). C(2/1) is cost involved when an observation actually drawn from $\boldsymbol{\pi}_1$ is incorrectly classified as $\boldsymbol{\pi}_2$.

Clearly, a good classification procedure is one which minimize in some sense or the cost of misclassification. Now , the expected cost of misclassification(ECM) is obtained by multiplying the off-diagonal entries in (11) by their probabilities of occurrence. Consequently a reasonable classification rule should has an ECM as small as possible. From the above the ECM may be defined as follows :

ECM= C(1/2) .P(misclassification into $\boldsymbol{\pi}_1$) + C(2/1) .P(misclassification into $\boldsymbol{\pi}_2$)

=C(1/2) .P(1/2). P_2 +C(2/1) . P(2/1) . P_1 → (12)

DEFINITION:

Expected (or average)cost of misclassification (ECM) is the sum of the products of costs of each misclassification multiplied by the probability of its occurrence. Its formula is given by (12).

THEOREM (Optional ECM regions or Bayes regions):

The regions R_1 & R_2 that minimize ECM are defined by the values of $\underset{\sim}{\mathbf{x}}$ for which the following inequalities hold.

$$R_1 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \geq \left(\frac{C(1/2)}{C(2/1)}\right) / \left(\frac{P_1}{P_2}\right) \quad \rightarrow (1)$$

(density ratio) $\geq$ (cost ratio)/(prior probability ratio)

$$R_2 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < \left(\frac{C(1/2)}{C(2/1)}\right) / \left(\frac{P_1}{P_2}\right) \quad \rightarrow (2)$$

PROOF:

We have the expected cost of misclassification (ECM) as

$$\text{ECM} = \text{C(1/2)} .\text{P(1/2)}\ P_2 + \text{C(2/1)}.\text{P(2/1)}\ P_1 \quad \rightarrow (3)$$

where C(1/2) = …….

C(2/1) = …….

P_1 = …….

P_2 = …….

P(1/2) = …….

P(2/1) = ...

But , we have

$$P(1/2) = \int_{R_1} f_2(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}}$$

$$P(2/1) = \int_{R_2} f_1(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} \quad \rightarrow (4)$$

using (4) in (3) we get

$$\text{ECM} = \text{C(1/2)}\ P_2 \int_{R_1} f_2(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} + \text{C(2/1)}\ P_1 \int_{R_2} f_1(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} \quad \rightarrow (5)$$

Noting that $\Omega = R_1 \cup R_2$ so that the total probability

$$1 = \int_{\Omega} f_1(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} = \int_{R_1} f_1(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} + \int_{R_2} f_1(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} \quad (\because R_1 \,\&\, R_2 \text{ are disjoint}) \rightarrow (6)$$

using (6) in (5),we get

$$\text{ECM} = \text{C}(1/2)\ P_2 \int_{R_1} f_2(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}} + \text{C}(2/1)\ P_1\left[1-\int_{R_1} f_1(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}}\right]$$

$$= \int_{R_1}[C(1/2)p_2 f_2(\underset{\sim}{\mathbf{x}}) - C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}})]d\underset{\sim}{\mathbf{x}} + C(2/1)p_1 \quad \rightarrow (7)$$

Now P_1, P_2,C(1/2) and C(2/1) are non-negative. In addition $f_1(\underset{\sim}{\mathbf{x}})$ & $f_2(\underset{\sim}{\mathbf{x}})$ are Non-negative for all $\underset{\sim}{\mathbf{x}}$ and are the only qualities in ECM that depend on $\underset{\sim}{\mathbf{x}}$. Therefore, minimization of ECM is equivalent to minimize the function

$$\int_{R_1}[C(1/2)p_2 f_2(\underset{\sim}{\mathbf{x}}) - C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}})]d\underset{\sim}{\mathbf{x}} \quad \rightarrow (8)$$

But, from the theory of integration (8) will be minimized is R_1 includes there values of $\underset{\sim}{\mathbf{x}}$ for which the integrand

$$\text{C}(1/2)\ P_2\ f_2(\underset{\sim}{\mathbf{x}}) - C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}}) \le 0 \quad \rightarrow (9)$$

and for all $\underset{\sim}{\mathbf{x}}$ those not included in R_1 or equivalently for all $\underset{\sim}{\mathbf{x}}$ those included in R_2

$$C(1/2)p_2 f_2(\underset{\sim}{\mathbf{x}}) - C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}}) > 0 \quad \rightarrow (10)$$

Thus from (9),

$$R_1 = \{\underset{\sim}{\mathbf{x}} / C(1/2)p_2 f_2(\underset{\sim}{\mathbf{x}}) - C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}}) \le 0\}$$

$$= \{\underset{\sim}{\mathbf{x}} / C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}}) \ge C(1/2)p_2 f_2(\underset{\sim}{\mathbf{x}})\}$$

$$= \left\{\underset{\sim}{\mathbf{x}} / \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \ge \left(\frac{C(1/2)}{C(2/1)}\right) / \left(\frac{p_1}{p_2}\right)\right\} \quad \rightarrow (11)$$

$(\because all f_1, f_2, p_1, p_2, C(1/2)\ \&\ C(2/1)$ are all positive)

similarly from (10),

$$R_2 = \left\{\underset{\sim}{\mathbf{x}} / \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < \left(\frac{C(1/2)}{C(2/1)}\right) / \left(\frac{p_1}{p_2}\right)\right\} \quad \rightarrow (12)$$

where (12) gives (2).

REMARK:

It is clear from (1) & (2) that the implementation of the minimum ECM

rules requires

(1). The ratio of p.d.f.'s is f_1 / f_2 is to be evaluated at a new observation $\underset{\sim}{\mathbf{x}}_0$.

(2). The cost ratio $\frac{C(1/2)}{C(2/1)}$

(3). The prior probability ratio $\frac{p_1}{p_2}$

The appearance of ratios in the definition of the optimal classification regions has significance as often it is much easier to specify the ratios than their component parts.

SPECIAL CASES OF MINIMUM "ECM" REGIONS

CASE (1):

(Equal prior probabilities i.e. $p_1 = p_2$ or $\frac{p_1}{p_2} = 1$)

In this case (1) & (2) become

$$R_1 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \geq \frac{C(1/2)}{C(2/1)};$$

$$R_2 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < \frac{C(1/2)}{C(2/1)}$$

CASE (2):

(Equal misclassification costs that is C(1/2)=C(2/1))

In this case $\frac{C(1/2)}{C(2/1)} = 1$ and therefore (1) & (2) become

$$R_1 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \geq \frac{p_2}{p_1};$$

$$R_2 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < \frac{p_2}{p_1}$$

CASE (3):

$$p_1 = p_2 \;\&\; C(1/2) = C(2/1)$$

In this case $\frac{p_1}{p_2} = 1 = \frac{C(1/2)}{C(2/1)}$ and therefore (1) & (2) become

$$R_1 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \geq 1;$$

$$R_2 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < 1$$

NOTE:

(1). When the prior probabilities are not known, they are often taken to be

equal.

(2). Similarly when the misclassification costs are unknown, they are often taken to be equal.

(3). If $\frac{C(1/2)}{C(2/1)} = \frac{p_1}{p_2}$ then $C(1/2)p_2 = C(2/1)p_1$ and hence

$$R_1 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \geq 1;$$

$$R_2 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < 1$$

OPTIMAL TOTAL PROBABILITY OF MISCLASSIFICATION (TPM) REGIONS :

Criteria other than the ECM can be used to derive "optimal" classification procedures. For example, one might ignore the costs of misclassification and

Choose R_1 & R_2 to minimize the total probability of misclassification (TPM).

TPM=P(misclassifying as $\boldsymbol{\pi_1}$ observation or misclassifying a $\boldsymbol{\pi_2}$ observation)

= P($\underset{\sim}{\mathbf{x}}$ comes from $\boldsymbol{\pi_1}$ and is misclassified)+

P($\underset{\sim}{\mathbf{x}}$ comes from $\boldsymbol{\pi_2}$ and is misclassified)

$$\Rightarrow TPM = P(\underset{\sim}{\mathbf{X}} \in R_2 / \boldsymbol{\pi_1}).P(\boldsymbol{\pi}_1) + P(\underset{\sim}{\mathbf{X}} \in R_1 / \boldsymbol{\pi}_2).P(\boldsymbol{\pi}_2)$$

$$= P_1 \int_{R_2} f_1(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} + P_2 \int_{R_1} f_2(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}}$$

$$= p_1 P(2/1) + p_2 P(1/2) \qquad \rightarrow (1)$$

But, when C(1/2)=C(2/1) (i.e. when misclassification costs are equal) we get from equation (12) of page 14,

$$ECM = C(1/2)[p_1 P(2/1) + p_2 P(1/2)] \qquad \rightarrow (2)$$

Now, from (1) & (2), it can be easily seen that minimizing (1) is equivalent to minimizing (2). In other words, minimizing TPM is equivalent to minimizing ECM with equal misclassification costs. Thus the optional TPM regions R_1 & R_2 are same as those given in case(2) of page 20. Thus

$$R_1 = \left\{ \underset{\sim}{\mathbf{x}} / \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \geq \frac{p_2}{p_1} \right\}$$

$$R_2 = \left\{ \underset{\sim}{\mathbf{x}} / \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < \frac{p_2}{p_1} \right\} \qquad \rightarrow (3)$$

ALLOCATING AN NEW OBSERVATION $\underset{\sim}{\mathbf{x}}_0$ BASED ON BAYE'S POSTERIOR PROBABILITIES

We can also allocate a new observation $\underset{\sim}{\mathbf{x}}_0$ to the population with the largest posterior probability $P(\boldsymbol{\pi}_i / \underset{\sim}{\mathbf{x}}_0)$. By Baye's rule, the "posterior" probabilities are

$P(\boldsymbol{\pi}_i / \underset{\sim}{\mathbf{x}}_0)$ = P($\boldsymbol{\pi}_1$ occurs and observe $\underset{\sim}{\mathbf{x}}_0$)/P(observe $\underset{\sim}{\mathbf{x}}_0$)

= P(observe $\underset{\sim}{\mathbf{x}}_0 / \boldsymbol{\pi}_1$).P($\boldsymbol{\pi}_1$)/

P(observe $\underset{\sim}{\mathbf{x}}_0 / \boldsymbol{\pi}_1$).P($\boldsymbol{\pi}_1$)+P(observe $\underset{\sim}{\mathbf{x}}_0 / \boldsymbol{\pi}_2$).P($\boldsymbol{\pi}_2$)

$$= \frac{f_1(\underset{\sim}{\mathbf{x}}_0).p_1}{f_1(\underset{\sim}{\mathbf{x}}_0).p_1 + f_2(\underset{\sim}{\mathbf{x}}_0).p_2}$$

$$= \frac{p_1 f_1(\underset{\sim}{\mathbf{x}}_0)}{p_1 f_1(\underset{\sim}{\mathbf{x}}_0) + p_2 f_2(\underset{\sim}{\mathbf{x}}_0)} \qquad \rightarrow (1)$$

$$P(\boldsymbol{\pi}_2 / \underset{\sim}{\mathbf{x}}_0) = 1 - P(\boldsymbol{\pi}_1 / \underset{\sim}{\mathbf{x}}_0)$$

$$= 1 - \frac{p_1 f_1(\underset{\sim}{\mathbf{x}}_0)}{p_1 f_1(\underset{\sim}{\mathbf{x}}_0) + p_2 f_2(\underset{\sim}{\mathbf{x}}_0)}$$

$$= \frac{p_2 f_2(\underset{\sim}{\mathbf{x}}_0)}{p_1 f_1(\underset{\sim}{\mathbf{x}}_0) + p_2 f_2(\underset{\sim}{\mathbf{x}}_0)} \qquad \rightarrow (2)$$

Now classify an observation $\underset{\sim}{\mathbf{x}}_0$ into $\boldsymbol{\pi}_1$ when

$$P(\boldsymbol{\pi}_1 / \underset{\sim}{\mathbf{x}}_0) > P(\boldsymbol{\pi}_2 / \underset{\sim}{\mathbf{x}}_0)$$

$$\Rightarrow p_1 f_1(\underset{\sim}{\mathbf{x}}_0) > p_2 f_2(\underset{\sim}{\mathbf{x}}_0)$$

($\because$ Numerators of (1) & (2) are equal)

$$\Rightarrow \frac{f_1(\underset{\sim}{\mathbf{x}}_0)}{f_2(\underset{\sim}{\mathbf{x}}_0)} > \frac{p_2}{p_1} \qquad \rightarrow (3)$$

Now from(3), it can be seen that allocating a new observation to a population based on Baye's posterior probabilities is same as optional TPM rule.

NOTE:

The above method is also equivalent to classify a new observation using optional ECM (Baye's method) rule when misclassification costs are equal.

CLASSIFICATION INTO ONE OF TWO KNOWN MULTIVARIATE NORMAL POPULATIONS (With common covariance matrix Σ)

Classification procedures based on normal populations predominate in statistical practice because of their simplicity and reasonably high efficiency across a wide variety of population models. We assume $f_1(\underset{\sim}{x})$ & $f_2(\underset{\sim}{x})$ are multivariate normal densities; the first with mean vector $\underset{\sim}{\mu}_1$,and the second with mean vector $\underset{\sim}{\mu}_2$ and both with common matrix $\boldsymbol{\Sigma}$. Now the p.d.f. of the two populations $\boldsymbol{\pi_1}$ & $\boldsymbol{\pi_2}$ is given by

$$f_i(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_i)} \quad \text{for i=1,2,....} \rightarrow (1)$$

The ratio of densities after simplification is

$$\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)+\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)}$$

$$= e^{-\frac{1}{2}\underset{\sim}{x}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{x}+\frac{1}{2}\underset{\sim}{x}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_1+\frac{1}{2}\underset{\sim}{\mu}_1'\boldsymbol{\Sigma}^{-1}\underset{\sim}{x}-\frac{1}{2}\underset{\sim}{\mu}_1'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_1+\frac{1}{2}\underset{\sim}{x}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{x}-\frac{1}{2}\underset{\sim}{x}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_2-\frac{1}{2}\underset{\sim}{\mu}_2'\boldsymbol{\Sigma}^{-1}\underset{\sim}{x}+\frac{1}{2}\underset{\sim}{\mu}_2'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_2}$$

$$\Rightarrow \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = e^{\underset{\sim}{\mu}_1'\boldsymbol{\Sigma}^{-1}\underset{\sim}{x}-\underset{\sim}{\mu}_2'\boldsymbol{\Sigma}^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_2)}$$

$$(\because \underset{\sim}{\mu}_i'\boldsymbol{\Sigma}^{-1}\underset{\sim}{x} = \underset{\sim}{x}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_i \ \ for \ \ i=1,2)$$

$$= e^{[(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\boldsymbol{\Sigma}^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)]} \rightarrow (2)$$

for $(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) = \underset{\sim}{\mu}_1'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_1'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_2-\underset{\sim}{\mu}_2'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_2$

$$= \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_2 \quad (\because \underset{\sim}{\mu}_1'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_2 = \underset{\sim}{\mu}_2'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mu}_1)$$

By minimum ECM classification rule, we have

$$R_1 : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq \left(\frac{C(1/2)}{C(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2 : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < \left(\frac{C(1/2)}{C(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

From (2) we have, after taking logarithms on both sides

$$R_1 : (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}\underset{\sim}{x} - \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}(\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}) \geq \log K$$

$$R_2 : (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}\underset{\sim}{x} - \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}(\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}) < \log K \qquad \rightarrow (3)$$

$$where \;\; K = \frac{C(1/2).p_2}{C(2/1)p_1}$$

The regions R_1 & R_2 given by (3) are called as minimum ECM regions for two normal populations.

NOTES:

(1). The first term of (3) viz

$$(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}\underset{\sim}{x} = \underset{\sim}{l}'\underset{\sim}{x}, \quad \text{where } \underset{\sim}{l} = \Sigma^{-1}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2}) \qquad \rightarrow (4)$$

is the well known Fisher (linear) discriminant function, which is actually obtained by Fisher with entirely different argument which we will discuss later.

(2). The minimum ECM classification rule for two normal populations is given by allocating $\underset{\sim}{x}_0$ to π_1 is

$$(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}\underset{\sim}{x}_0 - \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}(\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}) \geq \log K \qquad \rightarrow (5)$$

$$\text{where } K = \frac{C(1/2)p_2}{C(2/1)p_1}$$

otherwise, allocate $\underset{\sim}{x}_0$ to π_2

(3). The regions R_1 & R_2 given by equation (3) are called as best regions of classification.

(4). The regions R_1 & R_2 given by equation (3) are called as Baye's regions of classification.

(5). If the misclassification costs are equal i.e. C(1/2)=C(2/1), then the regions R_1 & R_2 given by (3) with $K = \frac{p_2}{P_1}$ are called as minimum(optional) TPM regions for two normal populations.

(6). The minimum or optional TPM classification rule for two normal populations is given by allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$ if

$$(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{x}}_0 - \frac{1}{2}(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\boldsymbol{\mu}}_1 + \underset{\sim}{\boldsymbol{\mu}}_2) \geq \log K \quad \rightarrow (6)$$

where $K = \frac{p_2}{p_1}$

(7). If misclassification costs are equal and prior probabilities are equal i.e. C(1/2)=C(2/1) and $p_1 = p_2$, then K=1 and consequently logK=0 and those from (3), we have

$$R_1 : (\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{x}} \geq \frac{1}{2}(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\boldsymbol{\mu}}_1 + \underset{\sim}{\boldsymbol{\mu}}_2)$$

$$\textit{and} \quad R_2 : (\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{x}} < \frac{1}{2}(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\boldsymbol{\mu}}_1 + \underset{\sim}{\boldsymbol{\mu}}_2) \quad \rightarrow (7)$$

In this case, the classification rule is allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$ if

$$(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{x}}_0 \geq \frac{1}{2}(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\boldsymbol{\mu}}_1 + \underset{\sim}{\boldsymbol{\mu}}_2) \quad \rightarrow (8)$$

allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$ otherwise.

This case may be used when both misclassification costs as well as prior probabilities are unknown.

(8). Let

$$U = (\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{X}} - \frac{1}{2}(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\boldsymbol{\mu}}_1 + \underset{\sim}{\boldsymbol{\mu}}_2)$$

$$= (\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}\left[\underset{\sim}{\mathbf{X}} - \frac{1}{2}(\underset{\sim}{\boldsymbol{\mu}}_1 + \underset{\sim}{\boldsymbol{\mu}}_2)\right]$$

Now let us obtain the distribution of U. since U is a linear combination of multivariate normal vector $\underset{\sim}{\mathbf{X}}$, U is distributed as univariate normal.

If $\underset{\sim}{\mathbf{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}_1, \boldsymbol{\Sigma})$

$$E(U) = \frac{1}{2}(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\Sigma^{-1}(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)$$

$$= \frac{\alpha}{2}, \text{ where } \alpha = (\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)$$

$$V(U) = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\operatorname{cov}(\underset{\sim}{X}, \underset{\sim}{X}')\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$$

$$= (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$$

$$= (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$$

$$= \alpha$$

Thus $U \sim N\left(\frac{\alpha}{2}, \alpha\right)$.

If $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}_2, \Sigma)$

In this case, we may show that $U \sim N\left(\frac{-\alpha}{2}, \alpha\right)$.

Here α is the mahalnobis squared distance D^2 between $\underset{\sim}{\mu}_1$ & $\underset{\sim}{\mu}_2$.

CLASSIFICATION INTO ONE OF TWO MULTIVARIATE NORMAL POPULATIONS WHEN THE PARAMETRERS ARE UNKNOWN

Suppose $\underset{\sim}{X}_{11}, \underset{\sim}{X}_{12}, \ldots \underset{\sim}{X}_{1n_1}$, be a random sample of size 'n_1', from population $\pi_1 : N(\underset{\sim}{\mu}_1, \Sigma)$ and let $\underset{\sim}{X}_{21}, \underset{\sim}{X}_{22}, \ldots, \underset{\sim}{X}_{2n_2}$ be a random sample of size 'n_2' from population $\pi_2 : N(\underset{\sim}{\mu}_2, \Sigma)$. Since $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2$ & Σ are unknown we replace them with their unbiased estimators viz.,

$$\underset{\sim}{\bar{X}}_1 = \frac{1}{n_1}\sum_{\alpha=1}^{n_1} \underset{\sim}{X}_{1\alpha}, \underset{\sim}{\bar{X}}_2 = \frac{1}{n_2}\sum_{\alpha=1}^{n_2} \underset{\sim}{X}_{2\alpha} \qquad \rightarrow (1)$$

and $$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \qquad \rightarrow (2)$$

where

$$S_1 = \frac{1}{n_1 - 1}\sum_{\alpha=1}^{n_1} (\underset{\sim}{X}_{1\alpha} - \underset{\sim}{\bar{X}}_1)(\underset{\sim}{X}_{1\alpha} - \underset{\sim}{\bar{X}}_1)'$$

$$S_2 = \frac{1}{n_2 - 1}\sum_{\alpha=1}^{n_2} (\underset{\sim}{X}_{2\alpha} - \underset{\sim}{\bar{X}}_2)(\underset{\sim}{X}_{2\alpha} - \underset{\sim}{\bar{X}}_2)' \qquad \rightarrow (3)$$

Now, the estimated (or sample) minimum ECM regions can be obtained from the above method replacing $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2$ & Σ with their unbiased estimators $\underset{\sim}{\bar{X}}_1, \underset{\sim}{\bar{X}}_2$ & S (given by (1) & (2)) respectively. They are form equations as follows :

$$\hat{R}_1 : (\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)' S^{-1} \underset{\sim}{X} - \frac{1}{2}(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)' S^{-1}(\underset{\sim}{\bar{X}}_1 + \underset{\sim}{\bar{X}}_2) \geq \log \left[c(1/2)p_2 / c(2/1)p_1\right] \quad \rightarrow (4)$$

$$\hat{R}_2 : (\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)' S^{-1} \underset{\sim}{X} - \frac{1}{2}(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)' S^{-1}(\underset{\sim}{\bar{X}}_1 + \underset{\sim}{\bar{X}}_2) < \log \left[c(1/2)p_2 / c(2/1)p_1\right] \quad \rightarrow (5)$$

from (4)&(5), the estimated sample minimum classification ECM rule for two normal populations is given by

Allocate $\underset{\sim}{X}_0$ to π_1 if

$$(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2) S^{-1} \underset{\sim}{X} - \frac{1}{2}(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)' S^{-1}(\underset{\sim}{\bar{X}}_1 + \underset{\sim}{\bar{X}}_2) \geq \log K \qquad \rightarrow (6)$$

where K= $\left[c(1/2)p_2 / c(2/1)p_1\right]$

Allocate $\underset{\sim}{X}_0$ to π_2 otherwise.

NOTE:

(1) The estimated or sample minimum TPM rule for two normal populations with unknown parameters can be obtained from (6) replacing K with $\left(p_2 / p_1\right)$.

(2) When $p_1 = p_2$ & $c(1/2) = c(2/1)$, the estimated or sample minimum ECM rule is equivalent to sample ML rule and is given by

Allocate $\underset{\sim}{x}_0$ to π_1 if

$$(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S^{-1} \underset{\sim}{x}_0 \geq \frac{1}{2}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S^{-1}(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2) \qquad \rightarrow (6)$$

Allocate $\underset{\sim}{x}_0$ to π_2 otherwise.

(3) The estimated minimum ECM rule or sample ML rule amounts to comparing the scalar variable (univariate normal variable)

$$y = \hat{\underset{\sim}{l}}' \underset{\sim}{x} \qquad \text{,where } \hat{\underset{\sim}{l}} = S^{-1}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2) \qquad \rightarrow (7)$$

evaluated at $\underset{\sim}{x}_0$ is

$$y_0 = \hat{\underset{\sim}{l}}' \underset{\sim}{x}_0$$

with the number

$$\hat{m} = \frac{1}{2}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S^{-1}(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2)$$

$$= \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \qquad \rightarrow (8)$$

where

$$\bar{y}_1 = (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)^{-1} S^{-1} \bar{\underset{\sim}{x}}_1 = \hat{\underset{\sim}{l}}' \bar{\underset{\sim}{x}}_1$$

$$\bar{y}_2 = \hat{\underset{\sim}{l}}' \bar{\underset{\sim}{x}}_2$$

Thus allocate $\underset{\sim}{x}_0$ to π_1 if

$$y_0 \geq \hat{m} \qquad \rightarrow (9)$$

otherwise allocate $\underset{\sim}{x}_0$ to π_2

That is, the estimated minimum ECM rule for two normal populations is

to creating two univariate normal populations for the y values by taking an appropriate linear combination of the observations from populations π_1 and π_2 and then assigning a new new observation $\underset{\sim}{x}_0$ to π_1 or π_2 depending upon whether $y_0 = \hat{\underset{\sim}{l}}' \underset{\sim}{x}_0$ falls to the right or left of the midpoint $\hat{m}$, between the two normal means $\bar{y}_1$ and $\bar{y}_2$.

(4) The linear function (7) is known as Fisher linear discriminant function. Which is obtained by Fisher with a different argument for separating two populations.

CLASSIFICATION OF NORMAL POPULATION WHEN $\Sigma_1 \neq \Sigma_2$

Here we have $\pi_1 : N(\underset{\sim}{\mu}_1, \Sigma_1)$ and $\pi_2 : N(\underset{\sim}{\mu}_2, \Sigma_2)$ when $\Sigma_1 \neq \Sigma_2$.

Let $f_1(\underset{\sim}{x})$ be the p.d.f. of π_1 and $f_2(\underset{\sim}{x})$ be the p.d.f. of π_2. Then on simplification,

$$\log\left[\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})}\right] = \log f_1(\underset{\sim}{x}) - \log f_2(\underset{\sim}{x})$$

$$= 1/2 \underset{\sim}{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\underset{\sim}{x} + (\underset{\sim}{\mu}_1' \Sigma_1^{-1} - \underset{\sim}{\mu}_2' \Sigma_2^{-1}) - \lambda \quad \rightarrow (1)$$

where $\lambda = 1/2 \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 1/2(\underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2) \quad \rightarrow (2)$

we have general formula for minimum ECM region and is given by

$$R_1 : \log[f_1(\underset{\sim}{x}) / f_2(\underset{\sim}{x})] \geq \log \text{ k} \quad \text{,where K=c(1/2)p}_2 / c(2/1)p_1$$

$$R_2 : \log[f_1(\underset{\sim}{x}) / f_2(\underset{\sim}{x})] < \log \text{ k} \qquad \rightarrow (3)$$

Now, the minimum ECM regions for classification of two normal populations when $\Sigma_1 \neq \Sigma_2$ is given by:

$$R_1 : -1/2 \underset{\sim}{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\underset{\sim}{x} + (\underset{\sim}{\mu}_1' \Sigma_1^{-1} - \underset{\sim}{\mu}_2' \Sigma_2^{-1})\underset{\sim}{x} - \lambda \geq \log \text{ k}$$

$$\text{R}_2 : -1/2 \underset{\sim}{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\underset{\sim}{x} + (\underset{\sim}{\mu}_1' \Sigma_1^{-1} - \underset{\sim}{\mu}_2' \Sigma_2^{-1})\underset{\sim}{x} - \lambda < \log \text{ k}$$

where λ & k are given as (2) & (3) → (4)

The allocation rule that minimizes the ECM is given by :

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$ if

$$-1/2\underset{\sim}{\mathbf{x}}_0'(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\underset{\sim}{\mathbf{x}}_0+(\underset{\sim}{\boldsymbol{\mu}}_1'\boldsymbol{\Sigma}_1^{-1}-\underset{\sim}{\boldsymbol{\mu}}_2'\boldsymbol{\Sigma}_2^{-1})\underset{\sim}{\mathbf{x}}_0-\lambda\geq\log\ \mathrm{k} \quad \rightarrow (5)$$

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$ otherwise.

In practice, the classification rule in (5) is implemented by substituting the sample quantities $\overline{\underset{\sim}{\mathbf{x}}}_1, \overline{\underset{\sim}{\mathbf{x}}}_2, S_1$ and S_2 for $\underset{\sim}{\boldsymbol{\mu}}_1, \underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ respectively.

QUADRATIC CLASSIFICATION RULE (NORMAL POPULATIONS WITH $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$)

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$ if

$$1/2\underset{\sim}{\mathbf{x}}_0'(S_1^{-1}-S_2^{-1})\underset{\sim}{\mathbf{x}}_0+(\overline{\underset{\sim}{\mathbf{x}}}_1'S_1^{-1}-\overline{\underset{\sim}{\mathbf{x}}}_2'S_2^{-1})\underset{\sim}{\mathbf{x}}_0-\hat{\lambda}\geq\log\ \mathrm{k} \quad \rightarrow (6)$$

allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$ otherwise.

$$\text{Where } \hat{\lambda}=1/2\log\left(\frac{|S_1|}{|S_2|}\right)+1/2(\overline{\underset{\sim}{\mathbf{x}}}_1'S_1^{-1}\overline{\underset{\sim}{\mathbf{x}}}_1-\overline{\underset{\sim}{\mathbf{x}}}_2'S_2^{-1}\overline{\underset{\sim}{\mathbf{x}}}_2) \quad \rightarrow (7)$$

NOTE:

(1). Minimum TPM rule or quadratic classification rule when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ is a special case of (6) when $K = p_2 / p_1$.

(2). If the misclassification costs are equal and prior probabilities are equal (i.e. $C(1/2) = C(2/1)$ & $p_1 = p_2$). Then the MC rule or QCR is obtained by taking K=1 or log K=0 in the rule (6).

FISHERS DISCRIMINANT FUNCTION-SEPARATION OF TWO POPULATIONS (NOT NECESSARY MULTIVARIATE NORMAL)

Fishers idea was to transform the multivariate observations $\underset{\sim}{\mathbf{x}}$'s to univariate observation y's such that the y's derived from population $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ were separated as much as possible . Fisher suggested taking linear combination of $\underset{\sim}{\mathbf{x}}$'s to create y's because they are simple function of $\underset{\sim}{\mathbf{x}}$ and are easily handled mathematically .

Fisher's approach does not assume that the populations are normal.

If does, however , implicitly assume the population covariance matrices are equal because a pooled estimate of the common covariance matrix is used.

Let $\underset{\sim}{\mathbf{x}}_{11},\underset{\sim}{\mathbf{x}}_{12},...,\underset{\sim}{\mathbf{x}}_{1n_1}$ be a random sample of size n_1 from population $\boldsymbol{\pi}_1$ and let $\underset{\sim}{\mathbf{x}}_{21},\underset{\sim}{\mathbf{x}}_{22},...,\underset{\sim}{\mathbf{x}}_{2n_2}$ be a random sample of size n_2 from population $\boldsymbol{\pi}_2$ .Now

$\overline{\underset{\sim}{\mathbf{x}}}_1$ be the mean of 1st sample

S_1 be the sample covariance matrix of 1st sample

$\overline{\underset{\sim}{\mathbf{x}}}_2$ be the mean of 2nd sample

S_2 be the sample covariance matrix of 2nd sample

Denote $S = \dfrac{(n-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$ → (1)

Which is a pooled sampled covariance matrix.

Now, Fisher's idea is as follows

Consider the linear combination

$y = \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}}$, when $\underset{\sim}{\mathbf{w}}$ is $|\mathbf{x}|$ vector of real number → (2)

using the linear transformation, the multivariate observation of 1st sample will be transformed into univariate observations given by

$y_{11},y_{12},...,y_{1n_1}$

when $y_{1i} = \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}}_{1i}, i = 1,2,...,n_1$

similarly the second sample $\underset{\sim}{\mathbf{x}}_{21},\underset{\sim}{\mathbf{x}}_{22},...,\underset{\sim}{\mathbf{x}}_{2n_2}$ will be transformed into

$y_{21},y_{22},...,y_{2n_2}$

when $y_{2i} = \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}}_{2i}, i = 1,2,...,n_2$

Now $\overline{y}_1 = \underset{\sim}{\mathbf{w}}'\overline{\underset{\sim}{\mathbf{x}}}_1$

$\overline{y}_2 = \underset{\sim}{\mathbf{w}}'\overline{\underset{\sim}{\mathbf{x}}}_2$

and $s_y^2 = \dfrac{\sum_{i=1}^{n_1}(y_{1i}-\overline{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\overline{y}_2)^2}{n_1+n_2-2}$ → (3)

consider

$\text{separation} = \dfrac{|\overline{y}_1 - \overline{y}_2|}{s_y}$ → (4)

Now, Fisher's idea is to select the linear combination $\underset{\sim}{\mathbf{w}}$ such that the separation given in (4) is maximum. In other words, the objective is to select the linear combination of $\underset{\sim}{\mathbf{x}}$ (i.e. $\underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}}$)

to achieve maximum separation between the sample means $\bar{y}_1$ & $\bar{y}_2$. Equation (4) may be written as

$$separartion^2 = \frac{\text{(squared distance between sample mean } \bar{y}_1 \text{ and } \bar{y}_2)}{\text{pooled sample variance of y}}$$

$$= \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

$$= \frac{\left[\underset{\sim}{w}'(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)\right]^2}{s_y^2} \qquad \text{(from (3))}$$

but $s_y^2 = \underset{\sim}{w}' S \underset{\sim}{w}$ (from (3) & (1))

$$\therefore \frac{\text{squared distance between } \bar{y}_1 \text{ \& } \bar{y}_2}{\text{pooled variance of y}} = \frac{(\underset{\sim}{w}'\underset{\sim}{d})^2}{\underset{\sim}{w}' S \underset{\sim}{w}} = \phi \text{ say} \qquad \rightarrow (5)$$

where $\underset{\sim}{d} = (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)$

Now, as per Fisher's idea, (5) has to be maximized w.r.t. $\underset{\sim}{w}$.

Which implies

$$\frac{\partial \phi}{\partial \underset{\sim}{w}} = \frac{(\underset{\sim}{w}' S \underset{\sim}{w}) \frac{\partial}{\partial \underset{\sim}{w}} (\underset{\sim}{w}'\underset{\sim}{d})^2 - (\underset{\sim}{w}'\underset{\sim}{d})^2 \frac{\partial}{\partial \underset{\sim}{w}} (\underset{\sim}{w}' S \underset{\sim}{w})}{\underset{\sim}{w}' S \underset{\sim}{w}} = 0$$

$$\Rightarrow (\underset{\sim}{w}' S \underset{\sim}{w}) 2 (\underset{\sim}{w}'\underset{\sim}{d}).\underset{\sim}{d} - 2(\underset{\sim}{w}'\underset{\sim}{d})^2 S \underset{\sim}{w} = 0$$

$$\Rightarrow (\underset{\sim}{w}' S \underset{\sim}{w})\underset{\sim}{d} - (\underset{\sim}{w}'\underset{\sim}{d}) S \underset{\sim}{w} = 0$$

$$\Rightarrow \underset{\sim}{w} = \frac{(\underset{\sim}{w}' S \underset{\sim}{w})}{(\underset{\sim}{w}'\underset{\sim}{d})} S^{-1}\underset{\sim}{d} \quad (\because \text{ S is positive defined matrix})$$

$$= CS^{-1}\underset{\sim}{d} \qquad \rightarrow (6)$$

where $C = \frac{\underset{\sim}{w}' S \underset{\sim}{w}}{\underset{\sim}{w}'\underset{\sim}{d}}$ and C is ratio of two scalars thus $\underset{\sim}{w}$ is a scalar multiplier of the vector $S^{-1}\underset{\sim}{d}$.

Using

$$\underset{\sim}{w} = CS^{-1}\underset{\sim}{d} \text{ in (5) we get}$$

$$\phi = \frac{C^2(\underset{\sim}{d}' S^{-1}\underset{\sim}{d})^2}{C^2 \underset{\sim}{d}' S^{-1} S S^{-1} \underset{\sim}{d}}$$

$$= \underset{\sim}{d}' S^{-1}\underset{\sim}{d} \qquad \rightarrow (7)$$

Now using $\underset{\sim}{w} = CS^{-1}\underset{\sim}{d}$ in (5) we get

$$\phi = \underset{\sim}{d}' S^{-1}\underset{\sim}{d} \qquad \rightarrow (8)$$

Thus, from (7) & (8), we can see that for either

$$\underset{\sim}{w} = CS^{-1}\underset{\sim}{d} \text{ or } \underset{\sim}{w} = S^{-1}\underset{\sim}{d}$$

the same ratio ϕ, we are setting . Thus ϕ will be maximized if we take

$$\underset{\sim}{\mathbf{w}} = S^{-1}\underset{\sim}{\mathbf{d}} = S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2) \qquad \rightarrow (9)$$

and the maximum value of ϕ is

$$\begin{aligned}\phi_m &= \underset{\sim}{\mathbf{d}}' S^{-1}\underset{\sim}{\mathbf{d}} \\ &= (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)' S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2) \\ &= D^2 \quad \text{(say)} \qquad \rightarrow (10)\end{aligned}$$

Now, the linear function

$$\begin{aligned}Y &= \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{X}} \\ &= (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)' S^{-1}\underset{\sim}{\mathbf{X}} \qquad \text{(from (2) \& (9))} \qquad \rightarrow (11)\end{aligned}$$

is called as Fisher's linear discriminant function . and the maximum ratio D^2 ,where D^2 given by (10),is called the sample squared distance or squared Mahalnobis distance between sample means $\underset{\sim}{\bar{\mathbf{x}}}_1$ and $\underset{\sim}{\bar{\mathbf{x}}}_2$.

The linear discriminant function given by (11) converts the two multivariate samples into two univariate samples such that the corresponding univariate sample means are sepearted as much as possible to the relative to pooled sample variance .

We can employ (11) as a classification device as given below.

AN ALLOCATION RULE BASED ON FISHER'S DISCRIMINANT FUNCTION :

We have the Fisher's linear discriminant function

$$y = \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}}, \qquad \text{where } \underset{\sim}{\mathbf{w}} = S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2) \qquad \rightarrow (1)$$

Let 'm' be the midpoint between $\bar{y}_1$ and $\bar{y}_2$ and is given by

$$\begin{aligned}m &= (\bar{y}_1 + \bar{y}_2)/2 \\ &= 1/2\ (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)' S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 + \underset{\sim}{\bar{\mathbf{x}}}_2) \qquad \rightarrow (2)\end{aligned}$$

Now , the allocation rule or classification rule based on Fisher's discriminant function is as follows:

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$,if

$$y_0 = (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)' S^{-1}\underset{\sim}{\mathbf{x}}_0 \geq m \text{ or } y_0 - m \geq 0$$

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$,if

$$y_0 < m \text{ or } y_0 - m < 0 \qquad \rightarrow (3)$$

NOTE:

(1). If $\boldsymbol{\pi}_1 \sim \underset{\sim}{\boldsymbol{\mu}}_1, \boldsymbol{\Sigma}$ and $\boldsymbol{\pi}_2 \sim \underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}$ then the Mahalnobis distance between $\underset{\sim}{\boldsymbol{\mu}}_1$ and $\underset{\sim}{\boldsymbol{\mu}}_2$ is

denoted by $\Delta_{\underset{\sim}{\mu}_1,\underset{\sim}{\mu}_2}$ and is given by

$$\Delta^2_{\underset{\sim}{\mu}_1,\underset{\sim}{\mu}_2} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$$

(2). $\Delta^2_{\underset{\sim}{x},\underset{\sim}{\mu}} = (\underset{\sim}{x} - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu})$

(3). Mahalnobis D^2 test statistic to test separation between π_1 *and* π_2 (or)

$H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$ vs $H_1 : \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$

suppose $\pi_1 : N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $\pi_2 : N_p(\underset{\sim}{\mu}_2, \Sigma)$ $\underset{\sim}{\bar{x}}_1, S_1$ are the sample mean and sample covariance matrix of a sample drawn from π_1 and $\underset{\sim}{\bar{x}}_2, S_2$ are …π_2.

Now Mahalnobis D^2 test statistic is given by

$$D^2 = (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S^{-1} (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)$$

where $S = \dfrac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$

under $H_0 : \left(\dfrac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}\right)\left(\dfrac{n_1 n_2}{n_1 + n_2}\right) D^2 \sim F_{p, n_1 + n_2 - p - 1}$

which can be used as for testing the significant difference $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$. If H_0 is rejected , we can conclude that the separation between the two populations π_1 *and* π_2 is significant.

(4). Two sample T^2 and Mahalnobis D^2 are closely associated as

$$T^2 = \left(\frac{n_1 n_2}{n_1 + n_2}\right) D^2$$

(5). In case of two normal populations with common covariance matrix, Fisher's method is corresponds to a particular case of minimum ECM rule with equal prior probabilities and equal costs of TPM rule with equal prior probabilities . Further, it is same as ML rule .

(6). The expression in minimum ECM rule for two multivariate normal populations $w = (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)' S^{-1} (\underset{\sim}{x} - 1/2(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2))$ is frequently called Anderson's classification.

(7). Fisher's method is also a special case of allocation rule based on Bayeson posterior probabilities when the prior probabilities p_1 and p_2 are same for the case of two multivariate normal populations.

CLASSIFICATION WITH SEVERAL POPULATIONS

THE MINIMUM "TPM" METHOD:

Let $\boldsymbol{\pi_1,\pi_2,...,\pi_g}$ are g populations. Let $f_i(\underset{\sim}{\mathbf{x}})$ be the density associated with population $\boldsymbol{\pi}_i, i$=l,2.…..g. Let

p_i=*the* prior probability of population $\boldsymbol{\pi}_i, i$=l,2,...g $\rightarrow$(1)

Now, for a given new observation $\underset{\sim}{\mathbf{x}}_0$, we have to allocate it to one of the populations $\boldsymbol{\pi_1,\pi_2,...,\pi_g}$. If is equivalent to, if the total sample space Ω is divided into $R_1, R_2, ..., R_g$ disjoint regions, and is the new observation $\underset{\sim}{\mathbf{x}}_0$ falls in region R_k, then $\underset{\sim}{\mathbf{x}}_0$ will be allocated to the population $\boldsymbol{\pi}_k$. Now, if $\underset{\sim}{\mathbf{x}}_0$ is actually drawn from population $\boldsymbol{\pi}_i$ and if it is allocated to $\boldsymbol{\pi}_k$ (i≠K) then we say that this is misclassification.

Thus, if R_k be the set of $\underset{\sim}{\mathbf{x}}'s$ classified as $\boldsymbol{\pi}_i$ and let

$$P(k/i)=P(\text{classify } \underset{\sim}{\mathbf{x}} \text{ in to } \boldsymbol{\pi}_k/\pi_i)$$
$$=\int_{R_k} f_i(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}} \qquad \rightarrow (2)$$

The conditional probability of misclassifying an $\underset{\sim}{\mathbf{x}}$ from $\boldsymbol{\pi_1}$ int*o* $\boldsymbol{\pi_2}$ **or** $\boldsymbol{\pi_3}$ **or... or** $\boldsymbol{\pi_g}$ is given by

$$CPM(1) = P(2/1)+P(3/1)+\ldots+P(g/1)$$
$$=\sum_{i=2}^{g} P(i/1) \qquad \rightarrow (3)$$

In a similar manner, we can obtain the conditional probabilities of misclassification CPM(2),CPM(3),…,CPM(g).

Multiplying each conditional CMP by its prior probability and summing gives the totals probability of misclassification(TPM).

Thus total probability of misclassification can be obtained as follows:

TPM=P(misclassification an observation $\underset{\sim}{\mathbf{x}}$)

=P(misclassification $\underset{\sim}{\mathbf{x}}$ from $\boldsymbol{\pi}_1$). $P(\boldsymbol{\pi}_1)$+…

+P(misclassification $\underset{\sim}{\mathbf{x}}$ from $\boldsymbol{\pi}_g$). $P(\boldsymbol{\pi}_g)$

$$= \sum_{i=1}^{g} CPM(i) P_i \qquad \text{where } p_i = P(\boldsymbol{\pi}_i)$$

$$= \sum_{i=1}^{g} P_i \left(\sum_{k \neq i=1}^{g} P(k/i) \right) \quad \text{(from (3))}$$

$$= \sum_{i=1}^{g} P_i \left(\sum_{k \neq i=1}^{g} \int_{R_k} f_i(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} \right) \qquad \rightarrow (4)$$

Determining optimal classification procedures based on TPM amounts choosing the mutually exclusive regions $R_1, R_2, ..., R_g$ such that (4) is minimum.

THE MINIMUM "ECM" METHOD:

The expected cost of misclassification (ECM) can be obtained as follows:

ECM=Expected cost of misclassifying an observation $\underset{\sim}{\mathbf{x}}$

=(conditional expected cost of misclassifying an $\underset{\sim}{\mathbf{x}}$ form $\boldsymbol{\pi}_1$) $P(\boldsymbol{\pi}_1)$

+…+(conditional expected cost of misclassifying an $\underset{\sim}{\mathbf{x}}$ from $\boldsymbol{\pi}_g$). $P(\boldsymbol{\pi}_g)$.

$$= p_1.ECM(1) + p_2.ECM(2) + ... + p_g.ECM(g)$$

$$= \sum_{i=1}^{g} p_i.ECM(i) \qquad \rightarrow (5)$$

where ECM(i)=conditional expected cost of misclassifying an $\underset{\sim}{\mathbf{x}}$ from $\boldsymbol{\pi}_i$

$$= \sum_{\substack{k=1 \\ k \neq i}}^{g} (\text{conditional expected cost of misclassifying an } \underset{\sim}{\mathbf{x}} \text{ from } \boldsymbol{\pi}_i \text{ into } \boldsymbol{\pi}_k)$$

$$= \sum_{\substack{k=1 \\ k \neq i}}^{g} c(k/i) \text{ (conditional probability of misclassifying an } \underset{\sim}{x} \text{ from } \boldsymbol{\pi}_i \text{ into } \boldsymbol{\pi}_k)$$

$$ECM(i) = \sum_{\substack{K=1 \\ K \neq i}}^{g} c(k/i) p(k/i) \qquad \rightarrow (6)$$

where $c(k/i)$=the cost of misclassifying $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ when actually it belongs to $\boldsymbol{\pi}_i (i \neq k)$

and $p(k/i)$ is as given by (2).

(Here it may be noted $c(i/i) = 0$, the misclassification cost for correct classification)

using (6) in (5), we get

$$ECM = \sum_{i=1}^{g} p_i \left[\sum_{\substack{K=1 \\ K \neq i}}^{g} c(k/i)\, p(k/i) \right]$$

$$= \sum_{i=1}^{g} p_i \left[\sum_{\substack{K=1 \\ K \neq i}}^{g} c(k/i) \int_{R_k} f_i(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} \right] \qquad \rightarrow (7)$$

Thus (7) gives the expected cost of misclassification , Now, an optional classification procedure obtained by minimizing ECM (given by (7)) for the choice of $R_1, R_2, ..., R_g$ is called minimum ECM method.

The choice of mutually exclusive regions $R_1, R_2, ..., R_g$ for when (7) is minimum are called minimum ECM regions.

THEOREM (MINIMUM "ECM" CLASSIFICATION RULE)

The classification regions that minimize the

$$ECM = \sum_{i=1}^{g} p_i \left[\sum_{\substack{K=1 \\ K \neq i}}^{g} c(k/i) \int_{R_k} f_i(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} \right]$$

are defined by allocating $\underset{\sim}{\mathbf{x}}$ to that population $\boldsymbol{\pi}_k$, k=1,2,...g

for which

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i c(k/i) f_i(\underset{\sim}{\mathbf{x}}) \qquad \rightarrow (8)$$

is smallest. If a tie occurs $\underset{\sim}{\mathbf{x}}$ can be assigned to any of the tied populations.

Minimum ECM classification rule with equal misclassification costs

(OR)

Minimum TPM classification rule:

Suppose all the misclassification costs are equal , in which case the minimum ECM rule reduces to minimum TPM rule. Thus in this case minimum ECM classification rule becomes (from (8))

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if

$$\sum_{\substack{i=1 \\ K \neq i}}^{g} p_i f_i(\underset{\sim}{\mathbf{x}}) \qquad \rightarrow (9)$$

is smallest. If tie occurs $\underset{\sim}{\mathbf{x}}$ can be assigned to any of the tied populations.

Now (9) will be smallest when the omitted term $p_k f_k(\underset{\sim}{\mathbf{x}})$ is largest. Consequently, the minimum ECM rule with equal misclassification costs (OR) minimum TPM rule is as follows:

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if

$$p_k f_k(\underset{\sim}{\mathbf{x}}) > p_i f_i(\underset{\sim}{\mathbf{x}}) \quad \forall i \neq k \qquad \rightarrow (10)$$

or equivalently, Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if

$$\log(p_k f_k(\underset{\sim}{\mathbf{x}})) > \log(p_i f_i(\underset{\sim}{\mathbf{x}})) \quad \forall i \neq k \qquad \rightarrow (11)$$

NOTES:

(1). It is interesting to note that the classification rule (10) is identical to the one that maximizes the posterior probability,

$$p(\pi_k / \underset{\sim}{\mathbf{x}}) = p(\underset{\sim}{\mathbf{x}} \text{ comes from } \pi_k \text{ given that } \underset{\sim}{\mathbf{x}} \text{ was observed})$$

$$= \frac{p_k f_k(\underset{\sim}{\mathbf{x}}))}{\sum_{i=1}^{g} p_i f_i(\underset{\sim}{\mathbf{x}})} = \frac{(prior) \times (likelihood)}{\sum (prior) \times (likelihood)}$$

Thus rule (10) or rule (11) is identical with the rule obtained based on Bayes posterior Probabilities.

(2). On case if prior probabilities are equal, rule (10) (or rule (11)) reduces to

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if

$$f_k(\underset{\sim}{\mathbf{x}}) > f_i(\underset{\sim}{\mathbf{x}}) \quad \forall i \neq k \qquad \rightarrow (12)$$

or equivalently Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if

$$\log f_k(\underset{\sim}{\mathbf{x}}) > \log f_i(\underset{\sim}{\mathbf{x}}) \quad \forall i \neq k \qquad \rightarrow (13)$$

The rule (12) or (13) is called as ML rule which is a special case of minimum TPM rule as well as minimum ECM rule.

(3). Generally , the minimum ECM rules have these components a) prior probabilities
b) misclassification costs
c) density functions

These components must be specified or estimated before the rules can be implemented.

CLASSIFICATION WITH NORMAL POPULATIONS

We have ‘g’ multivariate normal populations

$\boldsymbol{\pi}_1 : N_p(\underset{\sim}{\mu}_1, \boldsymbol{\Sigma}_1)$
$\boldsymbol{\pi}_2 : N_p(\underset{\sim}{\mu}_2, \boldsymbol{\Sigma}_2)$
$\vdots$
$\boldsymbol{\pi}_g : N_p(\underset{\sim}{\mu}_g, \boldsymbol{\Sigma}_g)$

In this case

$$f_i(\underset{\sim}{\mathbf{x}}) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_i^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)} \qquad \text{for i=1,2,....,g}$$

$$\log f_i(\underset{\sim}{\mathbf{x}}) = -\left(\frac{p}{2}\right)\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}((\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_i^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i)) \qquad \rightarrow (14)$$

CASE (1): UNEQUAL $\boldsymbol{\Sigma}_i$

In this case the minimum TPM rule (or minimum ECM rule with equal misclassification costs) is as follows:

From (14) we have

$$\log(p_k f_k(\underset{\sim}{\mathbf{x}})) = \log p_k - \frac{p}{2}\log(2\boldsymbol{\pi}) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_k\right| - \frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)'\Sigma_i^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)$$

Now the above rule becomes (from (11))

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if

$$\log(p_k f_k(\underset{\sim}{\mathbf{x}})) = \max_i \log(p_i f_i(\underset{\sim}{\mathbf{x}}))$$

$$= \log p_k - \frac{p}{2}\log(2\boldsymbol{\pi}) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_k\right| - \frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_i^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i) \qquad \rightarrow (15)$$

The constant $p/2\ \log(2\boldsymbol{\pi})$ can be ignored in (15) since it is same for all populations. We there fore define the quadratic discrimination score for i^{th} population is

$$d_i^Q(\underset{\sim}{\mathbf{x}}) = \log p_i - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_i^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i) \quad \text{for i=1,2,...,g} \qquad \rightarrow (16)$$

The quadratic score $d_i^Q(\underset{\sim}{\mathbf{x}})$ is composed contributions from the generalized variance $\left|\boldsymbol{\Sigma}_i\right|$, the prior probability p_i, and Mahalnobis (or statistical) squared distance between $\underset{\sim}{\mathbf{x}}$ and population mean μ_i.

Using discriminant scores the classification rule (15) becomes

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$

$$\text{The quadratic score } d_k^Q(\underset{\sim}{\mathbf{x}}) = \max_i \{d_i^Q(\underset{\sim}{\mathbf{x}})\} \qquad \rightarrow (17)$$

Where $d_i^Q(\underset{\sim}{\mathbf{x}})$ is given by (16).

In practice, the $\underset{\sim}{\boldsymbol{\mu}}_i$ and $\boldsymbol{\Sigma}_i$ are unknown and hence a training set of correctly classified observations is often available for the construction of estimates. The relevant sample quantities for population $\boldsymbol{\pi}_i$ are

$\overline{\underset{\sim}{\mathbf{x}}}_i$ = sample mean vector
S_i = sample covariance matrix and
n_i = sample size

Using the above estimation in (16) , we get the estimate of the quadratic discriminant score $\hat{d}_i^{Q}(\underset{\sim}{\mathbf{x}})$ as

$$\hat{d}_i^{Q}(\underset{\sim}{\mathbf{x}}) = \log p_i - \frac{1}{2}\log|S_i| - \frac{1}{2}(\underset{\sim}{\mathbf{x}} - \overline{\underset{\sim}{\mathbf{x}}}_i)' S_i^{-1}(\underset{\sim}{\mathbf{x}} - \overline{\underset{\sim}{\mathbf{x}}}_i) \qquad \rightarrow (18)$$

and the classification rule based on the sample is as follows:

ESTIMATED MINIMUM "TPM" RULE FOR SEVERAL NORMAL POPULATIONS- UNEQUAL $\boldsymbol{\Sigma}_i$

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if the quadratic score

$$\hat{d}_i^{Q}(\underset{\sim}{\mathbf{x}}) = \max_i \{\hat{d}_i^{Q}(\underset{\sim}{\mathbf{x}})\} \qquad \rightarrow (19)$$

where $\hat{d}_i^{Q}(\underset{\sim}{\mathbf{x}})$ is given by (18) for $i = 1, 2, ..., g$.

CASE-2:- $\boldsymbol{\Sigma}_i$'s ARE EQUAL

ESTIMATED MINIMUM "TPM" RULE (BAYES ALLOCATION RULE) NORMAL POPULATION –EQUAL $\boldsymbol{\Sigma}_i$:

In the case we have $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}(say)$ and hence the dicriminant score in (16) becomes

$$\hat{d}_i^{Q}(\underset{\sim}{\mathbf{x}}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\underset{\sim}{\mathbf{x}}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{x}} + \underset{\sim}{\boldsymbol{\mu}}_i'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{x}} - \frac{1}{2}\underset{\sim}{\boldsymbol{\mu}}_i'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\boldsymbol{\mu}}_i + \log p_i$$

The first two terms are same for $d_1^{Q}(\underset{\sim}{\mathbf{x}}), d_2^{Q}(\underset{\sim}{\mathbf{x}}), ... d_g^{Q}(\underset{\sim}{\mathbf{x}})$ and consequently, they can be ignored for allocatory purposes. Define the linear discriminant score

$$d_i(\underset{\sim}{\mathbf{x}}) = \underset{\sim}{\boldsymbol{\mu}}_i'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{x}} - \frac{1}{2}\underset{\sim}{\boldsymbol{\mu}}_i'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\boldsymbol{\mu}}_i + \log p_i \qquad \rightarrow (20)$$

An estimate of $d_i(\underset{\sim}{\mathbf{x}})$ viz, $\hat{d}_i(\underset{\sim}{\mathbf{x}})$ is based on the pooled estimate of $\boldsymbol{\Sigma}$,

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + ... + (n_g - 1)S_g}{n_1 + n_2 + ... + n_g - g} \qquad \rightarrow (21)$$

and is given by

$$\hat{d}_i(\underset{\sim}{\mathbf{x}}) = \overline{\underset{\sim}{\mathbf{x}}}_i' S^{-1} \underset{\sim}{\mathbf{x}} - \frac{1}{2} \overline{\underset{\sim}{\mathbf{x}}}_i' S^{-1} \underset{\sim}{\mathbf{x}} + \log p_i \qquad \rightarrow (22)$$

consequently, the estimated minimum TPM rule for equal covariance normal populations is as follows:

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if the linear discriminant score $\hat{d}_k(\underset{\sim}{\mathbf{x}}) = \max_i \{\hat{d}_i(\underset{\sim}{\mathbf{x}})\}$ $\rightarrow (23)$

Where $\hat{d}_i(\underset{\sim}{\mathbf{x}})$ is given by (22)

NOTE:-

(1). In the above minimum TPM rules , for any case , if $p_1 = p_2 = p_3 = ... = p_g = 1/g$, we may ignore those term $\log p_i$ is discriminant scores , as it is same for all discriminant scores. In this case the minimum TPM rule is reduced to ML rule in which case the allocation rules are same as above except ignoring $\log p_i$.

(2). Expression (20) is a convenient linear function of $\underset{\sim}{\mathbf{x}}$. An equivalent classifier for equal covariance matrices case can be obtained from (16) by ignoring the term $-\frac{1}{2}\log|\boldsymbol{\Sigma}|$ and is given by

$$-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i)' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i) + \log p_i$$

The classification rule with sample estimates instead for unknown populations quantities is given by Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$,if

$$-\frac{1}{2} D_k^2(\underset{\sim}{\mathbf{x}}) + \log p_k \text{ is largest} \quad \text{for k=1,2,…,g.} \qquad \rightarrow (24)$$

where $D_k^2(\underset{\sim}{\mathbf{x}}) = (\underset{\sim}{\mathbf{x}} - \overline{\underset{\sim}{\mathbf{x}}}_k)' S^{-1} (\underset{\sim}{\mathbf{x}} - \overline{\underset{\sim}{\mathbf{x}}}_k)$

is Mahalnobis squared distance between $\underset{\sim}{\mathbf{x}}$ and the sample mean $\overline{\underset{\sim}{\mathbf{x}}}_k$.

Thus , we see the rule (24) or equivalently rule (23) assigns $\underset{\sim}{\mathbf{x}}$ to the closest population (The distance in penalized by $\log p_i$).

(3). In note(2) , if we assume $p_1, p_2, p_3 ....., p_g$ are equal and hence allocation rule may be significant as follows:

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$, if $-\frac{1}{2} D_k^2(\underset{\sim}{\mathbf{x}})$ is largest

Or equivalently $D_k^2(\underset{\sim}{x})$ smallest $\rightarrow$ (25)

In other words , we are allocating $\underset{\sim}{x}$ to that population whose sample mean vector is closest to $\underset{\sim}{x}$. This rule is also called as ML classification rule.

FISHER'S METHOD FOR DISCRIMINATING AMONG SEVERAL POPULATIONS WHEN PARAMETERS ARE SPECIFIED

Fisher also proposed a several population extension of his discriminant method, which was discussed for the case of two populations. The motivation behind the Fisher discriminant analysis is the need to obtain a reasonable representation of the population that involves only a few linear combinations of the observations, such as $\underset{\sim}{l}_1'\underset{\sim}{x}, \underset{\sim}{l}_2'\underset{\sim}{x}$ and so on. His approach has several advantages and one is interested in separating several populations for

1) Visual inspection or

2) Graphical descriptive purposes.

It allows for the follows:-

1. Convenient representation of the g populations that reduce the dimension from a very large number of characteristics to a relatively few linear combinations. Of course, some information – needed for optimal classification- may be lost unless the population means lie completely in the lower dimensional space selected.

2. Plotting of the means of the first two or three linear combinations (discriminates). This helps display the relation ship and possible groupings of the populations.

3. Scatter plots of the sample values of the first two discriminates, which can indicate outliers or other abnormalities in the data.

The primary purpose of Fishers Discriminant analysis is to separate populations. However, it can also be used to classify a new observation into one of the populations. It is not necessary to assume that the g populations are multi variate normal. However we assume the population covariance matrices are equal and of full rank. That is $\Sigma_1 = \Sigma_2 = - - - = \Sigma_g = \Sigma$ .Thus, we have g populations with mean vectors $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots\ldots\underset{\sim}{\mu}_g$ and common covariance matrix $\sum$.

Let $\underset{\sim}{\bar{\mu}} = \frac{1}{g}\sum_{i=1}^{g}\underset{\sim}{\mu}_i$

and $B = \sum_{i=1}^{g}(\underset{\sim}{\mu}_i - \underset{\sim}{\bar{\mu}})(\underset{\sim}{\mu}_i - \underset{\sim}{\bar{\mu}})'$ --------**(1)**

we consider the linear combination $y = \underset{\sim}{l}\underset{\sim}{x}$

which has expected value

$E(y) = \underset{\sim}{l}'E(\underset{\sim}{X}/\pi_i) = \underset{\sim}{l}'\underset{\sim}{\mu}_i$ (for population π_i)

$= \mu_{iy}$ (say)

and variance $V(y) = \underset{\sim}{l}'cov(\underset{\sim}{X}, \underset{\sim}{X}')\underset{\sim}{l}$

$= \underset{\sim}{l}'\Sigma\underset{\sim}{l} = \sigma_y^2$ for all populations. --------**(2)**

we defuse the overall mean,

$$\bar{\mu}_y = \frac{1}{g}\sum_{i=1}^{g}\mu_{iy} = \frac{1}{g}\sum_{i=1}^{g}\underset{\sim}{l}'\underset{\sim}{\mu}_i$$

$$= \underset{\sim}{l}'\left(\frac{1}{g}\sum_{i=1}^{g}\underset{\sim}{\mu}_i\right)$$

$= \underset{\sim}{l}'\underset{\sim}{\bar{\mu}}$ (From (1)) -------**(3)** and form the ratio

$$\frac{\text{sum of squared distances from populations to over all mean of Y}}{\text{common population variance of Y}}$$

$$= \frac{\sum_{i=1}^{g}(\mu_{iy} - \bar{\mu}_y)^2}{\sigma_y^2}$$

$$= \frac{\sum_{i=1}^{g}(\underset{\sim}{l}'\underset{\sim}{\mu}_i - \underset{\sim}{l}'\underset{\sim}{\bar{\mu}})^2}{\underset{\sim}{l}'\Sigma\underset{\sim}{l}}$$

$$= \frac{\underset{\sim}{l}'\sum_{i=1}^{g}(\underset{\sim}{\mu}_i - \underset{\sim}{\bar{\mu}})(\underset{\sim}{\mu}_i - \underset{\sim}{\bar{\mu}})'\underset{\sim}{l}}{\underset{\sim}{l}'\Sigma\underset{\sim}{l}}$$

$$= \frac{\underset{\sim}{l}'B\underset{\sim}{l}}{\underset{\sim}{l}'\Sigma\underset{\sim}{l}}$$ (from (1))

Thus $$\frac{\sum_{i=1}^{g}(\mu_{iy}-\bar{\mu}_y)^2}{\sigma_y^2}=\frac{\underset{\sim}{l}'B\underset{\sim}{l}}{\underset{\sim}{l}'\sum\underset{\sim}{l}} \qquad \text{--------(4)}$$

The ratio (4) measures the variability between the groups of Y- values relative to the common variability within the groups. We can then choose $\underset{\sim}{l}$ to maximize the ratio (4) Thus if we write

$$\lambda=\frac{\underset{\sim}{l}'B\underset{\sim}{l}}{\underset{\sim}{l}'\sum\underset{\sim}{l}} \qquad \text{-------(5)}$$

Then we have to maximize (5) with respect to $\underset{\sim}{l}$ when implies

$$\frac{\partial\lambda}{\partial\underset{\sim}{l}}=0 \Rightarrow (\underset{\sim}{l}'\sum\underset{\sim}{l})\frac{\partial l'}{\partial\underset{\sim}{l}}B\underset{\sim}{l}-(\underset{\sim}{l}'B\underset{\sim}{l})\frac{\partial\underset{\sim}{l}'}{\partial\underset{\sim}{l}}\sum\underset{\sim}{l}=0$$

$$\Rightarrow (\underset{\sim}{l}'\sum\underset{\sim}{l})B\underset{\sim}{l}-(\underset{\sim}{l}'B\underset{\sim}{l})\sum\underset{\sim}{l}=0$$

$$\Rightarrow \quad B\underset{\sim}{l}-(\frac{\underset{\sim}{l}'B\underset{\sim}{l}}{\underset{\sim}{l}'\sum\underset{\sim}{l}})\,\Sigma\underset{\sim}{l}=0$$

$$\Rightarrow \sum^{-1}B\underset{\sim}{l}-(\frac{\underset{\sim}{l}'B\underset{\sim}{l}}{\underset{\sim}{l}'\sum\underset{\sim}{l}})\underset{\sim}{l}=0$$

$$\Rightarrow (\sum^{-1}B-\lambda I)\underset{\sim}{l}=0 \quad \text{(using (5))} \qquad \text{-------(6)}$$

Thus $\underset{\sim}{l}$ is the latent vector corresponding to a latent root λ of $\sum^{-1}B$. As, we are seeking for a $\underset{\sim}{l}$ which maximizes λ ,let λ_1 be the non zero largest latent root of $\sum^{-1}B$ and $\underset{\sim}{l}_1$ be the corresponding latent vector. Now, the linear combination, $Y_1=\underset{\sim}{l}_1'\underset{\sim}{X}$ is called Fisher's first linear discriminant .

Similarly if λ_2 is the next non Zero largest latent root of $\sum^{-1}B$ and $\underset{\sim}{l}_2$ correspondent latent vector then , $Y_2=\underset{\sim}{l}_2'\underset{\sim}{X}$ is Fisher's second linear discriminant.

Let $\lambda_1>\lambda_2>.....>\lambda_s>0$ denote the $s\le\min(g-1,p)$ non zero eigen values of $\sum^{-1}B$ and let $\underset{\sim}{l}_1,\underset{\sim}{l}_2,.....\underset{\sim}{l}_s$ be the corresponding latent vectors. Now, the linear combinations

$$Y_k=\underset{\sim}{l}_k'\underset{\sim}{X}(k\le s) \qquad \text{.....(7)}$$

is Fisher's k^{th} linear discriminant.

Fishers method for discriminating several populations when parameters are unknown

Fisher's sample linear discriminants:

In general, $\sum$ and the $\underset{\sim}{\mu_i'}s$ are unknown, but we have a training set consisting of correctly classified observations. Suppose the training set consist of a random sample of size n_i from population $\pi_i, i = 1,2,3.....g$

Let $\underset{\sim}{x_i'}$ be the mean vector and S_i be the covariance matrix of ith sample. Now denote the sample between groups matrix.

$$B_0 = \sum_{i=1}^{g} (\underset{\sim}{\bar{x}_i} - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}_i} - \underset{\sim}{\bar{x}})'$$

Where, $\underset{\sim}{\bar{x}} = \frac{1}{g}\sum_{i=1}^{g} \underset{\sim}{\bar{x}_i}$. **……(8)**

B_0 is an estimate of B

Also, an estimate of $\sum$ is based on the sample within groups matrix is

$$W = \sum_{i=1}^{g} (n_i - 1)S_i \qquad \textbf{…….(9)}$$

Consequently, $S_p = \frac{W}{(n-g)}, n = \sum_{i=1}^{g} n_i$. **……(10)**

is an estimate of $\sum$.

We consider the linear transformation,

$$y = \underset{\sim}{l'x} \qquad \textbf{……..(11)}$$

Under the linear transformation, (11) the given multi variate samples can be transformed into univariate samples whose means and variances are given by

Means : $\bar{y}_1, \bar{y}_{2,} .....\bar{y}_g$

Variances: $s_{y_1}^2, s_{y_2}^2, ...s_{y_g}^2$

We denote the overall sample as

$$\bar{y} = \frac{1}{g}\sum_{i=1}^{g} \bar{y}_i \qquad \textbf{……..(12)}$$

Now form the ratio,

$$\lambda = \frac{\text{sum of squared distances fro sample means to overall mean}}{\text{Total within samples variation}}$$

$$= \frac{\sum_{i=1}^{g}(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$$

$$= \frac{\sum_{i=1}^{g}(\underset{\sim}{l}'\underset{\sim}{\bar{x}}_i - \underset{\sim}{l}'\underset{\sim}{\bar{x}})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{l}'\underset{\sim}{x}_{ij} - \underset{\sim}{l}'\underset{\sim}{\bar{x}}_i)^2} \qquad \text{(from (11))}$$

$$= \frac{\underset{\sim}{l}'\sum_{i=1}^{g}(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})'\underset{\sim}{l}}{\underset{\sim}{l}'\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_j)(\underset{\sim}{x}_{ij} - \underset{\sim}{x}_i)'\underset{\sim}{l}}$$

$$= \frac{\underset{\sim}{l}'\sum_{i=1}^{g}(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}}_i)'\underset{\sim}{l}}{\underset{\sim}{l}'\sum_{i=1}^{g}(n_i - 1)S_i\underset{\sim}{l}}$$

(By using the definition of sample covariance matrix)

$$= \frac{\underset{\sim}{l}'B_0\underset{\sim}{l}}{\underset{\sim}{l}'W\underset{\sim}{l}} \qquad \text{(from (8) \& (9))} \qquad \textbf{…..(13)}$$

The ratio (13) measures the variability between the groups of g values relative to the total variability within the groups.

Now, Fisher suggested to choose $\underset{\sim}{l}$ such that λ given by (13) is maximum, Maximization of λ with respect to $\underset{\sim}{l}$ implies.

$$\frac{\partial\lambda}{\partial\underset{\sim}{l}} = 0 \Rightarrow (W^{-1}B_0 - \lambda I)\underset{\sim}{l} = 0 \qquad \textbf{…...(14)}$$

(See (6) of page 29 for derivation particulars)

Now, if we denote $\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_s > 0$ (where $s = \min(g-1, p)$) are s eigen values of (14) and let $\underset{\sim}{l}_1, \underset{\sim}{l}_2, ...\underset{\sim}{l}_s$ the corresponding eigen vectors, then

Fisher's K-th sample linear discriminant is given by

$$y_k = \underset{\sim}{l}_k' \underset{\sim}{x} (k \le s)$$

Thus, Fisher's sample linear discriminants are eigen vectors of $W^{-1}B_0$,

Where B_0 and W are as assigned in (8) & (9).

Note:

1) If sample means $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots\ldots\ldots \underset{\sim}{x}_g$ and sample covariance matrix $S_1, S_2, \ldots\ldots\ldots S_g$ are given , then B_0 and W can be completed using (8) and (9) respectively.

2) If raw samples from g populations are given, then B_0 and W can be computed as follows:

   First compute the individual sample covariance matrices $S_1, S_2, \ldots\ldots\ldots S_g$ from the given samples and then use (9) to compute W . Now, compute the sample covariance matrix S from the combined samples of g samples given by

$$S = \frac{1}{n-1} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}})' \quad \text{,when } n = \sum_{i=1}^{g} n_i \, .$$

   Now, B_0 can be computed from the following relationship.

$$(n-1)S = W + B_0 \quad .$$

3) It may be noted that the pooled sample covariance matrix S_p and combined sample covariance matrix S are connected by the

$$(n-1)S = W + B_0$$

Thus, if the individual sample covariance matrix $S_1, S_2 \ldots\ldots\ldots S_g$ and the combined sample covariance matrix S are given, then one can obtain W and B_0 can be obtained as follows

$$W = (n_1 - 1)S_1 + (n_2 - 1)S_2 + \ldots\ldots\ldots + (n_g - 1)Sg$$
$$B_0 = (n-1)S - W$$

Now, Fisher's discriminates can be constructed using the eigen vectors of $W^{-1}B_0$.

4) We know that $W^{-1}B_0$ is not a symmetric matrix. Many computer

Packages can compute eigen values and eigen vectors only for the symmetric matrices. However, the eigen vectors of $W^{-1}B_0$ can be computed as follows:

Suppose, λ is a ch root and $\underset{\sim}{l}$ is Ch. Vector of $W^{-1}B_0$,then we have

$$(W^{-1}B_0 - \lambda I)\underset{\sim}{l} = 0$$

The above equations may be rewritten as

$$(W^{-1}B_0 - \lambda W^{-\frac{1}{2}}W^{-\frac{1}{2}})\underset{\sim}{l} = 0$$
$$\Rightarrow (W^{-\frac{1}{2}}B_0 - \lambda W^{\frac{1}{2}})\underset{\sim}{l} = 0$$
$$\Rightarrow (W^{-\frac{1}{2}}B_0 W^{-\frac{1}{2}} - \lambda)W^{\frac{1}{2}}\underset{\sim}{l} = 0$$
$$\Rightarrow (W^{-\frac{1}{2}}B_0 W^{-\frac{1}{2}} - \lambda)\underset{\sim}{w} = 0$$

($W^{\frac{1}{2}}$ is square root of W and $W^{-\frac{1}{2}}$ is a inverse of $W^{\frac{1}{2}}$).

Where, $\underset{\sim}{w} = W^{\frac{1}{2}}\underset{\sim}{l}$ or $\underset{\sim}{l} = W^{-\frac{1}{2}}\underset{\sim}{w}$.

Thus if $\underset{\sim}{w}$ is latent vector of the matrix, $W^{-\frac{1}{2}}B_0 W^{-\frac{1}{2}}$. Corresponding to the latent root of λ, then latest vector $\underset{\sim}{l}$ of $W^{-\frac{1}{2}}B_0$ corresponding root λ may be obtained as $\underset{\sim}{l} = W^{-\frac{1}{2}}\underset{\sim}{w}$

For all practical purposes, for the construction of Fisher's discriminant functions we use the above method.

Classification of a new observation among several populations using Fisher's discriminants:-

Mainly, Fisher's discriminates were derived for the purpose of obtaining a low dimensional representation of the data that separate the populations as much as possible. Although they were derived from separatory considerations, the discriminates also provide the basis for a classification rule.

CLUSTER ANALYSIS

CONCEPT:

Multivariate methods deal with the analysis of data of more than two variables recorded from n sample objects selected from a specified population. Since the sample objects are selected from a specified population, the units are assumed to be homogeneous in respect of some characteristics. However, the values of different variables recorded from sample objects are not strictly uniform, though there should not be any systematic difference in the objects. In general, we expect some variations in the values of the variables, even if the sample objects are uniform in respect of some characters. For example, the income or the expenditure of middle class of people in a country are not exactly uniform, though they belong to the same class.

Again, the people of a country can be classified as rich, upper middle class, lower middle class and poor. For each class of people there may be common variable which influences the economic condition. For example, the income of a person depends on his education. This is true for every class of people. But their income or expenditure are not uniform. Therefore, there may be some systematic difference in values of the variables recorded from sample objects, there may be some similarities in the recorded observations of sample objects. Those sample objects which are similar in their recorded information may form a group. Dissimilar objects fall in different groups. In general, the objects that share similar characteristics are found together. In statistics, the search for relatively homogeneous objects is called cluster analysis.

The cluster analysis has wide applications in biology, medicine, agriculture, marketing, etc. The numerical taxonomy in the field of biology is used to classify the animals into class, order and families. Different species of plants have different characteristics. Therefore, plant specimens can be classify into homogeneous groups. In agriculture, the land fertility of a particular region may not be homogeneous for any type of crop. Then the pieces of land sharing similar fertility for a particular may be grouped together. The milk production of cows, even of the same type, may vary due to lactation period. Then the cows of the same lactation period may be grouped together. In

economics, the people of a city center may be grouped according to their socio-economic condition. In marketing, people can be grouped according to the similar buying habits. In medicine, the patients having similar disease may be clustered together.

Since similar objects form a cluster, all the sample points in any cluster will provide similar information about the population characteristics. Thus, for further analysis one may include one object from each cluster analysis is a data reduction technique in rows of the data matrix.

Let $X(n \times p)$ be a data matrix from a specified population. Let the values of the p variables observed from n sample objects be denoted by $X_1, X_2, ..., X_n$. The objective of the cluster analysis is to group these n vector of values into $n_1 (n_1 < n)$ vectors so that the elements in a group are homogeneous. Here the method of clustering is on the basis of one-sample observations. Let $X_{ij}[i = 1, 2, ..., n_j; j = 1, 2, ..., m]$ be the vector of values of p variables of i-th object in j-th sample. Here the objective of clustering is to form m_1 groups (m_1<m) of sample observations in different groups are heterogeneous.

From above discursion is is clear that the CA reduces the sample observations in size. It has similar property of other data reduction technique. Namely, PCA . this analysis has a similarity with DA in respect of classification of observations. But DA derives a rule for an allocating an object to its proper properties based on some prior information of the group membership of the object. Whereas, the CA identifies homogeneous groups or clusters.

There is no unified approach on what actually constitute a cluster. As per the definition what we have discussed above, a cluster constitutes with a similar object. Then, we need to decide on a measure of inter-object similarities. Also, a decision is needed to specify a procedure for forming the clusters, based on the chosen measure of similarity. The criterion of similarity in observations varies from researcher to researcher. However, the basic criterion is that the objects in a cluster should be closer to each other than to objects in other clusters. As a preliminary technique to identify the similarity of objects, one can use the diagram of sample objects. Let us consider that from each of 'n' sample object values of p variables are recorded. These values can be

represented a p-dimensional diagram. The values of each variable is plotted in each separate axis. If n sets of values are plotted in p-axes, a diagram will be formed. The cluster can be formed with those objects, which lie nearer in an area of the diagram but are dispersed from another area. The cluster can also be formed mathematically calculating distances among sample objects.

BASIC STEPS OF CA:

In CA, the sample objects are clustered on the basis of some characteristics. Therefore, to start with the analysis, a number of decisions must be made regarding the characteristics to be considered, the variables to be included in the analysis, the measurement of distance between objects and the criterion to group the objects.

The selection of variables for any CA is important, since the exclusion of important variables will be poor or misleading findings. For eg., if any marketing research the consumers are needed to be clustered, their tastes and habits and their economic capacities must be considered. Otherwise, the clustering of consumers will not be fruitful. The initial choice of variables determines the characteristics that can be used to identify subgroups.

After the selection of variables, the next important point to be considered is to measure the distance and similarity become objects. Two objects will be included in two separate groups, if their distance is maximum and they will be included in one group if they are close to each other. Therefore, one of the important steps in cluster analysis is to measure the distance among objects.

SIMILARITY MEASURES:

The measurement of similarity or distance is divided into two main parts. One of this is (a) distance – type measure, and another is (b) matching – type measure.

Distance type similarity measures:

Euclidean distance:

Let $\underset{\sim}{\mathbf{x}} = [x_1, x_2, ..., x_p]'$ and $\underset{\sim}{y} = [y_1, y_2, ..., y_p]'$ **are two p-dimensional observations or items or objects, then Euclidean distance between these two points is defined as**

$$d(\underset{\sim}{\mathbf{x}}, \underset{\sim}{\mathbf{y}}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2}$$
$$= \sqrt{(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\mathbf{y}})' (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\mathbf{y}})}$$

Statistical distance (Mahalonibis distance):

Let $\underset{\sim}{\mathbf{x}} = [x_1, x_2, ..., x_p]'$ and $\underset{\sim}{y} = [y_1, y_2, ..., y_p]'$ **are two p-dimensional observations or items or objects, then Statistical distance between these two points is defined as**

$$D(\underset{\sim}{\mathbf{x}}, \underset{\sim}{\mathbf{y}}) = \sqrt{(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\mathbf{y}})' \mathbf{A} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\mathbf{y}})}$$

Ordinarily, $\mathbf{A} = \mathbf{S}^{-1}$, where $\mathbf{S}$ contains the sample variances and covariances. However, without prior knowledge of the distinct groups, these sample quantities cannot be computed. For this reason Euclidean distance is often preferred for clustering.

Minkowski distance:

Let $\underset{\sim}{\mathbf{x}} = [x_1, x_2, ..., x_p]'$ and $\underset{\sim}{y} = [y_1, y_2, ..., y_p]'$ **are two p-dimensional observations or items or objects, then Minkowski distance between these two points is defined as**

$$d(\underset{\sim}{\mathbf{x}}, \underset{\sim}{\mathbf{y}}) = \left(\sum_{i=1}^{p} |x_i - y_i|^m \right)^{\frac{1}{m}}$$

Note 1: When m= 1 Minkowski distance is called as city block distance

Note 2: When m= 2 Minkowski distance is reduced to Euclidean distance thus Euclidean distance is special case of Minkowski distance.

Hierarchical clustering methods

Agglomerative hierarchical clustering algorithm:

The following are the steps in the agglomerative hierarchical clustering algorithm for groping *N* objects (items or variables) .

1. Starts with *N* clusters, each containing a single entity and an *N* X *N* symmetric matrix of distance (or similarites) $D=\{d_{ik}\}$.
2. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters *U* and *V* be d_{uv}.
3. Merge cluster *U* and *V*. Lable the newly formed cluster (*UV*). Update the enters in the distance matrix by (a) deleting the rows and columns corresponding to cluster *U* and *V* and (b) adding a row and column giving the distances between cluster (*UV*) and the remining clusters.
4. Repeat steps 2 and 3 a total of *N*-1 times.(All objects will be in a single cluster at termination of the algorithm,.) Record the identity of cluster that are merged and the levels(distances or similarities) at which the mergers take place.

Single Linkage (nearest neighbor) Clustering method:

In this method two objects with lowest distance are merged into a cluster. This cluster is the first one. At the second step, either a third object is merged to the first formed cluster, or the two closest unclustered objects are joined to form a second cluster. The decision rests on whether the distance from one of the unclustered objects to the first formed cluster is shorter than the distances between the two closest unclustered objects. The process continues until all objects belong to a single cluster. At any step, if a new cluster is formed and if the distance between the new cluster and the old one is shortest, then both the clusters are combined. **Since an object of a cluster cannot be split from the cluster, the distance between two clusters is equal to the distance between nearest objects**. After clustering, the objects in the clusters can be represented by a diagram. This diagram is known as Dendrogram.

The following are the steps in the single linkage clustering algorithm for groping *N* objects (items or variables) .

1. Starts with *N* clusters, each containing a single entity and an *N* X *N* symmetric matrix of distance (or similarites) $D=\{d_{ik}\}$.
2. Search the distance matrix for the smallest distance (other than zero) which gives the distance between nearest (most similar) pair of clusters. Let the distance between "most similar" clusters *U* and *V* be d_{UV}.
3. Merge clusters *U* and *V*. Label the newly formed cluster (*UV*). Update the enteries in the distance matrix by

   (a) deleting the rows and columns corresponding to cluster *U* and *V*

   and (b) adding a row and column giving the distances between the cluster (*UV*) and the remaining clusters. Here, the distance between the cluster (*UV*) and any other cluster *W* is computed by

   $$d_{(UV)W} = \min\{d_{UW}, d_{VW}\}$$

   In general, if we take two clusters, one is with *m* objects and another is with *n* objects, then we will get *'mn'* pairs of objects, one is from the first cluster and another is from the second cluster. Thus, we get totally *'mn'* distances. **Now the distance between these two clusters is defined as the minimum (smallest) of these *'mn'* distances.**

4. Repeat steps 2 and 3 a total of *N*-1 times.(All objects will be in a single cluster at termination of the algorithm,.) Record the identity of cluster that are merged and the levels(distances or similarities) at which the mergers take place.

Complete Linkage (farthest neighbor) Clustering method:

In this method two objects with lowest distance are merged into a cluster. This cluster is the first one. At the second step, either a third object is merged to the first

formed cluster, or the two closest unclustered objects are joined to form a second cluster. The decision rests on whether the distance from one of the unclustered objects to the first formed cluster is shorter than the distances between the two closest unclustered objects. The process continues until all objects belong to a single cluster. At any step, if a new cluster is formed and if the distance between the new cluster and the old one is shortest, then both the clusters are combined. **The distance between two clusters is defined as the distance between most distant pair of objects.** After clustering, the objects in the clusters can be represented by a diagram. This diagram is known as Dendrogram.

The following are the steps in the compete linkage clustering algorithm for groping *N* objects (items or variables) .

1. Starts with *N* clusters, each containing a single entity and an *N* X *N* symmetric matrix of distance (or similarites) $D=\{d_{ik}\}$.
2. Search the distance matrix for the smallest distance (other than zero) which gives the distance between nearest (most similar) pair of clusters. Let the distance between "most similar" clusters *U* and *V* be d_{UV}.
3. Merge clusters *U* and *V*. Label the newly formed cluster (*UV*). Update the enteries in the distance matrix by

   (a) deleting the rows and columns corresponding to cluster *U* and *V*

   and (b) adding a row and column giving the distances between the cluster (*UV*) and the remaining clusters. Here, the distance between the cluster (*UV*) and any other cluster *W* is computed by

   $$d_{(UV)W} = \max\{d_{UW}, d_{VW}\}$$

   In general, if we take two clusters, one is with *m* objects and another is with *n* objects, then we will get *'mn'* pairs of objects, one is from the first cluster and another is from the second cluster. Thus, we get totally *'mn'* distances. **Now the distance between these two clusters is defined as the maximum (largest) of these *'mn'* distances.**

4. Repeat steps 2 and 3 a total of *N*-1 times.(All objects will be in a single cluster at termination of the algorithm,.) Record the identity of cluster that are merged and the levels(distances or similarities) at which the mergers take place.

Average Linkage Clustering method:

In this method two objects with lowest distance are merged into a cluster. This cluster is the first one. At the second step, either a third object is merged to the first formed cluster, or the two closest unclustered objects are joined to form a second cluster. The decision rests on whether the distance from one of the unclustered objects to the first formed cluster is shorter than the distances between the two closest unclustered objects. The process continues until all objects belong to a single cluster. At any step, if a new cluster is formed and if the distance between the new cluster and the old one is shortest, then both the clusters are combined. **The distance between two clusters is defined as the average of the distances between all pairs of objects, one from the first cluster and another from the second cluster.** After clustering, the objects in the clusters can be represented by a diagram. This diagram is known as Dendrogram.

The following are the steps in the compete linkage clustering algorithm for groping *N* objects (items or variables) .

1. Starts with *N* clusters, each containing a single entity and an *N* X *N* symmetric matrix of distance (or similarites) $D=\{d_{ik}\}$.
2. Search the distance matrix for the smallest distance (other than zero) which gives the distance between nearest (most similar) pair of clusters. Let the distance between "most similar" clusters *U* and *V* be d_{UV}.
3. Merge clusters *U* and *V*. Label the newly formed cluster (*UV*). Update the enteries in the distance matrix by

   (a) deleting the rows and columns corresponding to cluster *U* and *V*

   and (b) adding a row and column giving the distances between the cluster

(UV) and the remaining clusters. Here, the distance between the cluster (UV) and any other cluster W is computed by

$$d_{(UV)W} = \left(d_{UW} + d_{VW}\right)/2$$

In general, if we take two clusters, one is with m objects and another is with n objects, then we will get *'mn'* pairs of objects, one is from the first cluster and another is from the second cluster. Thus, we get totally *'mn'* distances. **Now the distance between these two clusters is defined as the average (mean) of these *'mn'* distances.**

4. Repeat steps 2 and 3 a total of N-1 times.(All objects will be in a single cluster at termination of the algorithm,.) Record the identity of cluster that are merged and the levels(distances or similarities) at which the mergers take place.

Nonhierarchical clustering methods

Nonhierarchical clustering techniques are designed to groups items, rather than variables, into a collection of K clusters. The number of clusters, K, may either be specified in advance or determined as part of the clustering procedure. Because oa matrix of distances (similarities) does not have to be determined and the basic data do not have to be stored during the computer rum, nonhierarchical methods can be appled to much larger data sets than hierarchical techniques.

Nonhierarchical methods start from either (1) an initial partition of itms into groups or (2) an initial set of seed points, which will form the nuclei of clusters. Good choice for starting configurations should be free of overt biases. One way to star is to randomly select seed points from among the items or to randomly partition the items into initial groups.

In this section we discuss one of the more poplar nonhierarchical procedure, the K-means method.

K-means method:

MacQueen suggests the term K-means for describing his algorithm that assigns each item to the cluster having the nearest centroid (mean). In its simplest version, the process is composed of these three steps.

1. Partition the items into K initial clusters.
2. Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. (Distance is usually computed using Euclidean distance with

either standardized or unstandardized observations.) Recalculate the centriod for the cluster receiving the new item and for the cluster losing the item.

3. Repeat Step2 until now more reassingnments take place.

Rather than starting with a partition of all items into K preliminary groups in Step 1, we could specify K initial centroid (seed points) and then proceed to Step.

The final assignment of items to clusters will be, to some extent, dependent upon the initial partition or the initial selection of seed points. Experience suggests that most major changes in assignment occur with the first reallocation step.

UNIT-1

1. (2001)

(a) Define the generalized inverse of an $m \times n$ matrix and derive four important properties.

(b) How do you define the joint , conditional and marginal densities of vectors of random variables.

(OR)

(c) Explain and distinguish uncorrelated, orthogonal and independent random vectors.

(d) Define eigen values and eigen vectors. Establish their properties.

(20002)

(a) Define the generalized inverse of an $m \times n$ matrix and indicate a method of obtaining the generalized inverse of a square matrix whose determinant is zero.

(b) For a given random vectors X and Y, prove that

(1) cov(AX,BY)=A[cov(X,Y)] B' and

(2) cov(X,Y)= $E[XY'] - E(X)[E(Y)]'$.

(OR)

(c) Show that every matrix satisfies its own characteristic function.

(d) If X is a $p \times 1$ vector of random variables with mean vector, θ and Dispersion matrix, Σ show that $E[X'AX] = tr(A\Sigma) + \theta' A\theta$ where **A** is a symmetric matrix of order $p \times p$.

(2003)

(a) Define non-singular multivariate normal (MVN) distribution and obtain its characteristic function.

(b) Let $X \sim N_p(\mu, \Sigma)$. If $X'=(X^{(1)'}, X^{(2)'})$, where $X^{(1)'}$ has K-components and $X^{(2)'}$ has (p-k) components and further μ and Σ are partitioned appropriately. Then find the conditional distribution of $X^{(1)} / X^{(2)} = x^{(2)}$.

(OR)

(c) Let $X \sim N_p(\mu, \Sigma)$ and suppose that X is partitioned in the form $X=(X^{(1)}, X^{(2)})'$, where $X^{(1)} : (q \times 1)$ and $X^{(2)} : ((p-q) \times 1)$. Then $X^{(1)}$ and $X^{(2)}$ are statistically independent if and only if $\text{cov}(X^{(1)}, X^{(2)}) = \underset{\sim}{0}$.

(d) Obtain the maximum likelihood estimators of the parameters of a MVN distribution.

(2004)

(a) Define the multivariate normal distribution and obtain its characteristic function.

(b) Define :

(1) Marginal and

(2) Conditional distributions

Prove that the conditional distribution obtained from the multivariate normal distribution is normal.

(OR)

(c) Show that if X is $N_p(\mu, \xi)$, then the marginal distribution of any subset of X is multivariate normal with mean. Variances and covariances obtained by taking proper components of μ and ξ.

(d) Let μ and ξ be the parameters of a multivariate normal distribution. Assuming a random sample from this distribution, obtain the maximum Likelihood. Estimate for μ and ξ.

(2006)

(a) Define the characteristic function of a p-dimensional random variables. Establish its properties.

(b) Find the covariance matrix fore the two random variables X_1 and X_2 when their joint probability function, $P_{12}(x_1, x_2)$ is represented by the entries in the body of the following table:

x_1 \ x_2	0	1	$p_1(x_1)$
−1	0.24	0.06	0.3
0	0.16	0.14	0.3
1	0.40	0.00	0.4
$p_2(x_2)$	0.8	0.2	1

(OR)

(c) Obtain the maximum likelihood estimators of the parameters of a multivariate normal distribution.

(d) Let X be $N_3(\mu, \xi)$ with

$$\Sigma = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Are X_1 and X_2 independent? What about (X_1, X_2) and X_3?

(Model paper)

(a) Define the multivariate normal distribution and obtain its characteristic function.

(b) Prove that the conditional distribution obtained from the multivariate normal distribution is normal.

(OR)

(c) Show that if X is $N_p(\mu,\Sigma)$ then the marginal distribution of any subset of X is multivariate normal with mean ,variances and co-variances obtained by taking proper components of μ and Σ.

(d) Let μ and ξ be the parameters of a multivariate normal distribution . assuming a random sample from this distribution, obtain the maximum likelihood. Estimates for μ and Σ.

(Model paper)

(a) Define multivariate normal distribution and obtain its characteristic function.

(b) Define marginal and conditional distributions. Prove that the conditional distribution obtained from the multivariate normal distribution is also multivariate normal distribution.

(2007)

(a) Derive the density function of a p-variate normal distribution .

(b) Let X has a trivariate normal distribution with $E(X)=\tilde{0}$ and variance-covariance matrix $\Sigma=\begin{bmatrix} 1/2 & -1/2 & 1/2 \\ -1/2 & 1 & -1/2 \\ 1/2 & -1/2 & 1 \end{bmatrix}$. Find the conditional distribution of X_1 given $X_2=x_2$ and $X_3=x_3$.

(OR)

(c) Obtain the characteristic function of multivariate normal distribution .

(d) Derive the ML estimators of the mean vector and covariance matrix in the case of p-variate normal distribution.

2. (2001)

(a) X is a p-variate random vector having the distribution $N_p(\mu,\Sigma)$

where μ is $(p\times 1)$ and Σ is $(p\times p)$. X is partitioned as

$X=\begin{pmatrix} X_{(1)} \\ X_{(2)} \end{pmatrix}$ where X_1 is of order $(q\times 1)$. μ and Σ are also conformably partitioned. Derive (1) the marginal distribution of $X_{(1)}$ and (2) the conditional distribution of $X_{(1)}$ given $X_{(2)}=x_{(2)}$.

(b) Define the p-variate normal distribution with mean vector μ and dispersion matrix Σ. Derive three important properties of the multivariate normal distribution.

(OR)

(c) In the p-variate normal case, show that the sample mean vector and the sample covariance matrix are independently distributed.

(d) Derive the M.L. estimators of the mean vector and the covariance matrix in the multivariate normal distribution $N_p(\mu,\Sigma)$.

(2002)

(a) Define multivariate normal distribution and obtain its characteristic function.

(b) If **X** is $N_p(\mu,\Sigma)$, prove that $X'\Sigma^{-1}X$ has a chi-square distribution with P degrees of freedom.

(OR)

(c) State and prove a necessary and sufficient condition for $X'AX$ and $X'BX$ to be independently distributed, given that X is $N_p(\mu,\Sigma)$.

(d) Derive the sampling distribution of the sample correlation coefficient (γ) when the population correlation coefficient (ρ) is zero.

(2003)

(a) Define Hotelling's T^2 and derive its probability density function. Explain the uses of this distribution.

(b) Discuss a test procedure for testing the equality of mean vectors of two MVN populations, assuming equal dispersion matrices.

(OR)

(c) stating the assumptions clearly , discuss the problem of comparing several multivariate normal population means.

(e) Explain the following:

(1) Wilks lambda criterion.

(2) Need for simultaneous confidence intervals.

(2004)

(a) Derive the null distribution of Hotelling's T^2 statistic.

(b) Show that Hotelling's T^2 statistic can be used to test equality of means of corresponding variables in two multivariate normal populations having the same variance-covariance matrix.

(OR)

(c) Explain in detail the procedure of carrying out MANOVA of one way classification.

(d) Define simultaneous confidence regions and illustrate broad steps to obtain the same.

(2006)

(a) Derive the null distribution of Hotelling's T^2 statistic.

(b) Explain in detail the likelihood ratio principle.

(OR)

(c) Explain in detail the one way classification of multivariate data. Write the MANOVA table.

(d) Discuss the Behern-Fisher problem.

(Model paper)

(a) Derive the null distribution of Hotelling's T^2 statistic.

(b) Show that Hotelling's T^2 statistic can be used to test the equality of means of corresponding varianles in two multivariate normal populations having the same variances-covariance matrix.

(OR)

(c) Explain in detail the procedure of carrying out MANOVA of one way classification.

(d) Define simultaneous confidence regions and illustrate broad steps to obtain the same.

(Model paper)

(a) Derive the MLEs of the mean vector , μ and the variance-covariance matrix , Σ based on a random sample of size n draw from the normal population $N_p(\mu, \Sigma)$.

(b) Show that the sample mean vector and sample variance-covariance matrix obtained based on a random sample of size n from a normal populations $N_p(\mu, \Sigma)$ are independently distributed and also mention their distributions.

(2007)

(a) Define Hotelling's T^2 statistic. What is its distribution? What are the applications of T^2? Explain the relationship between Hotelling's T^2 and Mahalnobis D^2.

(b) Discuss a test procedure for testing the equality of mean vector of two multivariate normal populations having equal dispersion matrix.

(OR)

(c) Explain MANOVA one way classification with an example.

(d) Explain in detail the likelihood ratio principle.

3. (a) How do you test for the goodness of fit of a linear model?

(b) What is path analysis? What purposes are served by it?

(OR)

(c) What is multicollinearity? Discuss the methods of overcoming multicollinearity .

(d) Define autocorrelation . What are the consequences of the presence of autocorrelation?

(2002)

(a) State and prove Gauss-Markoff theorem for a standard general linear model , $\underset{n\times 1}{Y} = \underset{n\times k}{X}\ \underset{\mathrm{K}\times 1}{\beta} + \underset{\mathrm{n}\times 1}{\mathrm{u}}$.

(b) Explain the concept of hetroskedasticity and describe a test procedure for testing the same.

(OR)

(c) Explain the concept of multicolinearity and indicate the consequences of OLS estimation of the parameters in the presences of collinearity.

(d) Explain the concept of auto-correlation and describe Durbin-Watson d-statistic for detecting the same.

(2003)

(a) What do you understand by " dimension reduction"? describe a technique that is used for this purpose.

(b) Define:

(1) Canonical variables and

(2) Canonical correlation

Explain how do you estimate canonical correlation under the normality assumption.

(OR)

(c) What are various properties of principal components? Show that principal Components are all uncorrelated.

(d) Prove that canonical correlations are invariant under non-singular transformations and any function of the variance-covariance matrix that is invariant is a function of the canonical correlations.

(2004)

(a) Describe the principal components analysis.

(b) What are canonical variates and canonical correlations? How do you compute them?

(OR)

(c) Define principal components and discuss their use in statistical analysis. If $N_p(\mu, \xi)$ then explain how you would compute the various principals components.

(d) Show that the canonical correlations are invariant under non-singular linear transformations of $X^{(1)}, X^{(2)}$ variables of the form $\underset{(p\times p)}{C}\ \underset{(p\times 1)}{\mathrm{X}^{(1)}}$ and $\underset{(q\times q)}{D}\ \underset{(p\times 1)}{\mathrm{X}^{(2)}}$.

(2006)

(a) What is principal component analysis? How are they useful?

(b) Suppose the random variables X_1, X_2 and X_3 have the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

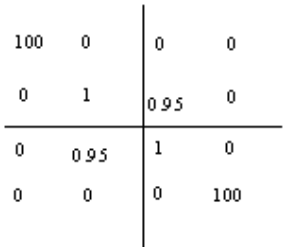
The corresponding eigen values are 5.8285,2.00,0.1716. obtain the three principal components and their variances.

(OR)

(c) What are canonical variables? Explain how these are useful in the analysis

of multivariate data.

(d) For the following covariance matrix



Obtain the first pair of canonical variables and the canonical correlation between them.

(Model paper)

(a) Describe the principle component analysis.

(b) What are canonical variates and canonical correlations? How do you compute them.

(OR)

(c) Define principle components and discuss their uses in statistical analysis . If $N_p(\mu,\Sigma)$ then explain how you would compute the various principle components.

(d) Show that canonical correlations are invariant under non-singular transformations of $X^{(1)}, X^{(2)}$ variables of the form $\underset{(p\times p)}{C}\ \underset{(p\times 1)}{X^{(1)}}\ \underset{(q\times q)}{D}\ \underset{(p\times 1)}{X^{(2)}}$.

(Model paper)

(a) Derive the null distribution of Hotelling's T^2 statistic.

(b) If X is $N_p(\mu_i,\Sigma)$, prove that $X'\Sigma^{-1}X$ has a chi-square distribution with p degrees of freedom .

(2007)

(a) What do you mean by dimension reduction? Discribe a technique that is used for this purpose.

(b) Let the random variables X_1, X_2, X_3 have the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Obtain the three principal components.

(OR)

(c) Define canonical variables and canonical correlations. Explain how do you estimate canonical correlation under the normality assumption.

(d) Show that the canonical correlations are invariant under non-singular linear transformations of $X^{(1)}, X^{(2)}$ variables of the form $C_{p\times p}, X^{(1)}_{p\times 1}$ and $D_{q\times q} X^{(2)}_{q\times 1}$.

4. (2001)

(a) Explain principal component analysis. Give the uses of principal components in the factor analysis.

(b) Describe cluster analysis and explain its uses.

(OR)

(c) Explain canonical correlations. Show that the multiple correlation coefficient is a special case of canonical correlation coefficient. How will you obtain the latter?

(d) What are canonical variables? Explain how these are useful in the analysis of multivariate data.

(2002)

(a) Define principal components and explain the procedure of obtaining them.

(b) What are canonical variates and canonical correlations? How do you compute them?

(OR)

(c) distinguish between factor analysis and principal component analysis.

(d) Given two sets of variates $X^{(1)}$ & $X^{(2)}$, show that the canonical correlation are invariant under non-singular linear transformations of the form

$Y^{(1)} = AX^{(1)}, Y^{(2)} = BX^{(2)}$ where A and B are non-singular matrices.

(2003)

(a) Let $f_i(x) \sim N_p(\mu_i, \Sigma), i = 1, 2.....$. Derive the optimal classification rule.

Also obtain the probabilities of misclassification.

(b) Describe the following hierarchical clustering methods:

(1) single linkage method.(SLINK)

(2) complete linkage method (CLINK)

(OR)

(c) Discuss Fisher's method for discrimination among several populations.

(d) Explain the concept of clustering. What is a Dendogram? The following of sample correlations for five stocks:

$$D = \begin{bmatrix} 1 & & & & \\ .58 & 1 & & & \\ .51 & .60 & 1 & & \\ .39 & .39 & .44 & 1 & \\ .46 & .32 & .43 & .52 & 1 \end{bmatrix}$$

Treating the sample correlation coefficients as similarity measures, cluster analyze the stocks using the nearest neighbour method and also draw the dendogram.

(2004)

(a) Explain the problem of classification and derive fishers linear discriminant function.

(b) Explain the importance of cluster analysis and describe of hierarchial clustering.

(OR)

(c) Explain the problem of classification into one of the two known multivariate normal populations.

(d) Describe the method of classification of an individual into one of several p-variate normal populations having a common dispersion matrix. ξ where all the parameters are known.

(2006)

(a) What is meant by discriminant analysis? Establish the relationship between Fishers discriminant function and Mahalnobi's D^2 statistic.

(b) Explain the importance of cluster analysis and describe the Hierarchial method of clustering.

(OR)

(c) Explain Fiher's method of discriminating between two multivariate populations.

(d) A researcher has enough data available to estimate the density functions $f_1(\underset{\sim}{\mathbf{x}})$ and $f_2(\underset{\sim}{\mathbf{x}})$ associated with populations $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_2}$ respectively. Suppose C(2/1)=5 units,C(1/2)=10 units and it is known that about 20% of all objects (for which the measurement $\underset{\sim}{\mathbf{x}}$ can be recorded) belonging $\boldsymbol{\pi_2}$.

Suppose the density function evaluated at a new observation $\underset{\sim}{\mathbf{x}}_0$ give $f_1(\underset{\sim}{\mathbf{x}}_0)$=0.3 and $f_2(\underset{\sim}{\mathbf{x}}_0)$=0.4. Do you classify the new observation as $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_2}$.

(Model paper)

(a) Explain the problem of classification and derive Fishers linear discriminant function.

(b) Let $f_i(X) \sim N_p(\mu_i, \Sigma)$,i=1,2,…. . Derive the optimal classification rule also obtain the probabilities of misclassification.

(c) Explain the problem of classification into one of the known multivariate normal populations.

(d) Describe the method of classification of an individual in to one of several p-variate populations having a common dispersion matrix Σ where all the parameters are known .

(Model paper)

(a) Stating the assumptions clearly , discuss the problem of comparing several multivariate normal population means.

(b) Explain the following:

(1) Likelihood ratio test

(2) Need for simultaneous confidence intervals.

(2007)

(a) Explain the problem of classification and derive Fisher's linear discriminant function.

(b) Explain the problem of classification into one of the two known multivariate normal populations.

(OR)

(c) What are the standards of good classification? Describe a linear discriminant function as a tool for classification.

(d) Describe the method of classification of an individual into one of several p-variate normal populations having a common dispersion matrix when the parameters are known.

5. (2001)

(a) What is the problem of classification and what are the standards of good classification?

(b) Discuss the problem of classification into one of two known multivariate normal populations.

(OR)

(c) Explain the concept of discriminant function and discuss,with an example , its use in discriminatory analysis.

(d)When do you apply cluster analysis ? Discuss single linkage method with an example.

(2002)

(a) Explain the problem of classification into one of the two known multivariate normal populations.

(b) Distinguish between cluster analysis and discriminant analysis.

(OR)

(c) Describe the method of classification of an individual into one of several p-variate normal populations having a common dispersion matrix, Σ where all the parameters are known.

(d) Explain the importance of cluster analysis and describe the method of Hierarchical clustering.