Major Based Elective - II

Multivariate Analysis

18KP2SELS2

Unit - I

Aspects of multivariate Analysis, Applications of multivariate techniques - Some basics of matrix and vector algebra - Mean vectors and covariance matrix - Generalised variance - Multivariate normal distribution - Multivariate normal density and its Properties.

Unit - II

Hotelling T^2 - Statistics : Introduction - derivation and its distribution - Uses of T^2 statistics, Properties of T^2, Wishart distribution - definition and properties only.

Unit - III

Principle Components : Introduction - Population Principle Components - Summering Sample Variation by Principle Components - Graphing the Principle Components.

Unit - IV

Factor analysis and Inference for Structured Covariance matrices, Orthogonal factor model - methods of estimation - Factor rotation - factor scores.

Unit - V

Discrimination and classification - Separation and classification of two Populations - classification with two multivariate normal populations Evalulating classification functions - Fisher's discriminant function - Fisher's method for discriminating among several populations.

UNIT-I

Ascepts of Multivariate Analysis :

Multi Variate analysis :

A data included Simultaneously measurement on many variables This body of methodology this called multivariance analysis.

Objectives of Multivariate analysis :

(i) The data reduction (or) stucture the statisfy value information In this hoper thus that is from maker the interepation Specked.

(ii) The Starting and grouping :

The groups of "Similar" objects or Variables are Created the bascal occur measure, the based occur measure, the characteristic, rules for classifying object in to will define groups required.

(iii) The Investigation depend among Variables :

The major of the relationship of interest or the Variables mutually independent or one or more variables depends on the others.

(iv) Prediction : The relationship between Variables must be determined for the purpose of prediting the values of one or more on the basis of observation on the other variables.

(v) Hypothesis : Specific Statistical hypothesis formulated interms of the Parameter of multi Variate assumption or to reinforce prior Convictions.

2

Application of Multivariate Technique :

(i) Data reduction of Simplification :

Using data can several variable data related to cancer patient responses to radio theraphy a simple measure of patient response to constructs. A matrix of tactic similarities was developed from aggregate data derived from professional which propessional mediator judge the facties they are in resolving dispates was determined.

(ii) Sorting and grouping :

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

(iii) Prediction :

The association between the tests course and several high school performance variable and several college for performance variables were used to develop prepiitase success in college.

(IV) Hypothesis Testing :

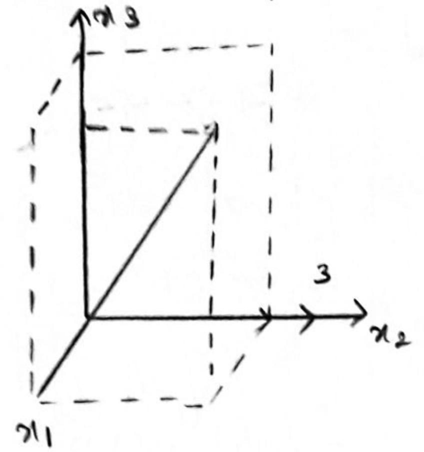
Several population - related variables where measured to determine whether measured to determine whether throughly constant throughout of wear (or) whether ther was a noticable difference between the weak days and week ends. The use of methods in diverse field

Some Basic of Matrix and Vector Algebra Vector :

An away on a real numbers $x_1, x_2 \cdots x_n$ is called n Vectors,

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ (or) } (X') = [X_1, X_2 \cdots X_n]$$

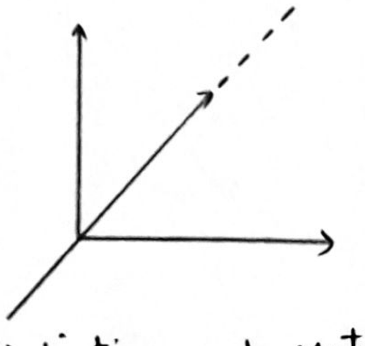
Where the Prime denotes the Operation of transposing a column to a row:



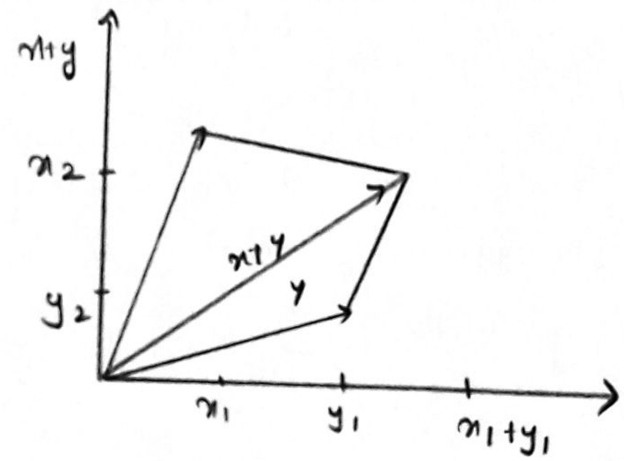
We define the vector C_x as,

$$C_x = \begin{bmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_n \end{bmatrix}$$

That is C_x is the Vector obtained by multiplying each element of X by C.



Scalar multiplication and vector addition,



Two vectors both direction and length In $n=2$ dimensions We consider the vector $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$X+Y = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1+y_1 \\ x_2+y_2 \\ \vdots \\ x_n+y_n \end{bmatrix}$$

The length of X Written L_x is defined to be,

$$L_x = \sqrt{x_1^2 + x_2^2}$$

The length of a Vector $x' = (x_1, x_2 \cdots x_n)$ with n Components is defined by, $L_x = \sqrt{x_1^2 + x_2^2 + x_n^2}$ (2-1)

Multiplication of a Vector x by a Scalar changes the lengths,

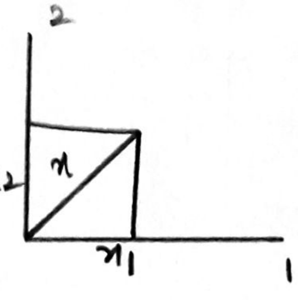
$$L_{cx} = \sqrt{c_1 x_1^2 + c_2^2 x_2^2 + \cdots c^2 x_n}$$
$$= |c| \sqrt{x_1^2 + x_2^2 + \cdots x_n^2}$$
$$= |c| \, L_x.$$

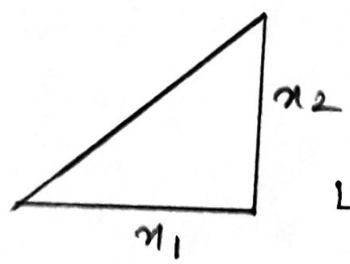
Multiplication by c does not change the direction of the Vector X if $c > 0$.

However a negative where of C Create a vector with a direction opposite that of X from.

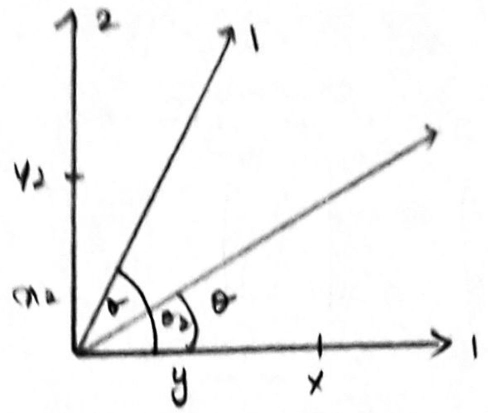
$$L_{cx} = |c| L_x$$

If is clear that x is if $|c| > 1$ and $0 < k|k| < 1$ Chooshing $c = L_x^{-1}$ we obtain the Unit Vector $L_y^{-1} x$. length L and lies in the direction of X. $L_x = \sqrt{x_1^2 + x_2^2}$.





Length of $x = \sqrt{x_1^2 + y_2^2}$.



$$\cos\theta = \frac{x_1y_1 + x_2y_2}{L_x L_y}.$$

The angle θ between $x' [x_1, x_2]$ and $y' = [y_1, y_2]$.
A second geometrical concept angles consider the vector in the angle θ between the θ can be represented as the difference between the angles θ_1 and θ_2 from by the two vectors and the first co-ordinates axis.

$$\cos(\theta_1) = \frac{x_1}{L_x}, \quad \cos(\theta_2) = \frac{y_1}{L_y}, \quad \sin(\theta_1) = \frac{x_2}{L_x}, \quad \sin(\theta_2) = \frac{y_2}{L_y}$$

and,

$$\cos(\theta) = \cos(\theta_2 - \theta_1)$$

$$= (\cos\theta_2 \cos\theta_1 + \sin\theta_2) \sin\theta_1$$ the angle θ between the two vector,

$x' = (x_1, x_2)$ and $y' = (y_1, y_2)$ is specified by

$$\cos(\theta) = \cos(\theta_2 - \theta_1) = (y_1/L_y)(x_1/L_x) + (y_2/L_y)(x_2/L_x)$$

$$= \frac{x_1y_1 + x_2y_2}{L_x L_y}$$

$$x'y = x_1y_1 + x_2y_2$$

With this definition and equation,

$L_x = \sqrt{x'x}$

$$\cos(\theta) = \frac{x'y}{L_x L_y} = \frac{x'y}{\sqrt{x'x}\sqrt{y'y}}$$

Sin θ $\cos(90°) = \cos(270°) = 0$

and $\cos(\theta) = 0$ and only if $x'y = 0$ x and y are prependicular

When $x'y = 0$

$$x'y = x_1y_1 + x_2y_2 + \cdots + x_ny_n$$

Using the inner Product we have the actual extension of length and angle to vectors.

L_x = length of $x = \sqrt{x'x}$

$$\cos(\theta) = \frac{x'y}{L_xL_y} = x'y / \sqrt{x'x}\sqrt{y'y}$$

Since again $\cos(\theta) = 0$ only if $x'y = 0$ we say that x and y are prerpendicular when $x'y = 0$

Generalized Variance:

One multivariate analougue of the Variance σ^2 of a univariate distribution is the co-variance matrix Σ. Another multivariate analogue is the Scalar $|\Sigma|$ which is called the generalized variance.

The Scalar $|\Sigma|$ in the multivariate analogue is called the generalized variance of the multivariate distribution.

The generalised Variance to the Sample of vectors, $X_1, X_2 \cdots X_n$ is,

$$|S| = \left| \frac{1}{N-1} \sum_{\alpha=1}^{N} (X_\alpha - \bar{X})(X_\alpha - \bar{X})' \right|$$

Multi Normal distriubtion :

Derivation of Multi Variate Normal Distribution :

The Multi Variate Normal distribution is generalization of the Univariate normal distribution.

If $X \sim N(\mu, \sigma^2)$ then the density function of this Variable is,

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-1/2 \left(\frac{x-\mu}{\sigma}\right)^2} \rightarrow (1)$$

The term $\left[\frac{(x-\mu)}{\sigma}\right]^2$ measure the Square distance from x to μ in Standard deviation (σ) units. This can be written as,

$$\left[\frac{x-\mu}{\sigma}\right]^2 = (x-\mu)\frac{1}{\sigma^2}(x-\mu)'$$

$b = \mu$, and $\alpha = 1/\sigma^2$

$$\left[\frac{x-\mu}{\sigma}\right]^2 = (x-b)\,\alpha\,(x-b)$$

If we replace the random Variable 'x' by a Vector X.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}$$

and α be the 'f' we define matrix A

$$\alpha = A \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & & & \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}$$

Then the term,

$$(x-b)\,\alpha\,(x-b) = (x-b)\,A\,(x-b)'$$

Now ① can be written as,

$$f(x) = f(x_1, x_2 \cdots x_p)$$

$$f(x) = K e^{-1/2}\,(x-b)\,A\,(x-b)' \rightarrow (2)$$

In the Variate of univariate case the Constant is called the Normalizing factor $\left[1/\sqrt{2\pi}\,\sigma\right]$

Now we have to determine 't' we quatity 'K' Such that the integral value of the function ② over p - dimensional equal to space must be unity.

<u>Additive Property :</u>

Let X [with ϕ components] be distributed according to N (M, Σ) then $Y = CX$ is distributed according to N $(CM, C\Sigma C')$ for C is non-singular.

Proof :

If X is given that $X \sim N_p(\mu, \Sigma)$ then its mgf is,

$$M_X(T) = e^{T'\mu + T'\Sigma T/2}$$

$$= E\left[e^{T'x}\right]$$

We have,

$$Y = CX$$

$$M_Y(T) = E\left[e^{T'y}\right] \qquad (\because L' = T'C)$$

$$= E\left[e^{L'x}\right]$$

If is given that $X \sim N_p(\mu, \Sigma)$

$$M_Y(T) = e^{L'\mu + \frac{L'\Sigma T}{2}}$$

Substituting for L,

$$M_Y(T) = e^{T'C\mu + \frac{T'(C\Sigma C)T}{2}}$$

$$= e^{T(C\mu) + \frac{T'(C\Sigma C')T}{2}}$$

$$\Rightarrow Y \sim N_p(C\mu, C\Sigma C')$$

Hence its proved.

UNIT - II

Hotelling T^2 - Statistics :

Definition :

If $X_1, X_2 \cdots X_N$ be N - independent observation from a P - variate normal (μ, Σ) there the T^2 - Statistics is defined as,

$$\frac{T^2}{N-1} = N(\bar{X}-\mu)' A^{-1} (\bar{X}-\mu)$$

Where, N is a sample size,

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

and

$$A = \sum_{i=1}^{N} (X_i - \bar{X})' (X_i - \bar{X}).$$

In the Univariates case of $x_1, x_2 \ldots x_n$ are n independent observation from $N(\mu, \sigma^2)$ to test the hypothesis $H_0 : \mu = \mu_0$. We use the statistic agains, $H_1 = \mu \neq \mu_0$ (two sided).

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (\because \sigma \text{ is unknown})$$

Squaring on both sides,

$$t^2 = \frac{(\bar{X}-\mu)^2 n}{S^2}$$

$$t^2 = n(\bar{x}-\mu)' \left(\frac{1}{S^2}\right) (\bar{x}-\mu)$$

$$t^2 = n(\bar{x}-\mu)' (S^2)^{-1} (\bar{x}-\mu)$$

If we replace the squares by the multivariate analogue, We get,

$$t^2 = n(\bar{x}-M)' S^{-1} (\bar{x}-M)$$

$$T^2 = N (\bar{X}-M)' S' (\bar{X}-M)$$

Where, N is a Sample Size,

$$\bar{X} = \frac{\Sigma X_i}{N}$$

$$S = \frac{1}{N-1} A$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})' (x_i - \bar{x})$$

Replacing S by A

$$T^2 = N (\bar{X}-M)' A^{-1} (\bar{x}-\mu)$$

$$T^2 = N (\bar{x}-M)' A' (\bar{x}-\mu).$$

Derivation of Distriubtion of Hotelling T^2 Statistic:

We Know that,

$$\frac{T^2}{N-1} = N (\bar{X}-M_0)' A' (\bar{X}-M_0)$$

$$= \sqrt{N} (\bar{X}-M_0)^* S' \sqrt{N} (\bar{x}-M_0)$$

$$T^2 = \sqrt{N} (\bar{x}-M_0)^* S' \sqrt{N} (\bar{X}-M_0)$$

$$T^2 = y' S^{-1} y.$$

Where,

$$y = \sqrt{N} (\bar{x}-\mu_0)$$

$$E(Y) = \sqrt{N} (\bar{X}-M_0) = \sqrt{N} (M-M_0)$$

$$V(Y) = \Sigma = N \cdot \frac{\Sigma}{N}$$

$$Y \sim N (\sqrt{N} (M-M_0) \Sigma)$$

(i.e) $Y \sim N(\gamma, \Sigma)$ Where, $E(Y) = \sqrt{N} (M-M_0)$.

To derive the distribution of T^2 under non- Case to obtain to distribution of T^2 assuming.

$$Y \sim N(\mathcal{L}, \Sigma)$$

$$\text{Let } (N-1)S = \sum_{i=1}^{N-1} (x_i \times x_i)' \qquad (\because S = \sum_{i=1}^{N-1} x_i \times x_i)$$

Where $Z_i \sim N(0, \Sigma)$ and are i.i.d

Let 'D' be a non- singular matrix

$$\Rightarrow D^* \Sigma D' = x$$

Define, $Y^* = DY$, $S^* = DSD'$

$$\Rightarrow Y = Y^* D^{-1} \text{ and } S = D^{-1} S^* (D)^{-1}$$

$$\therefore Y^* \sim N(\mathcal{L}^* x) \qquad (\because \mathcal{L}^* = D\mathcal{L})$$

$$T^2 = YS^{-1}Y$$

$$T^2 = (Y^*)' (S^*)' (Y^*)$$

Consider,

$$(N-1)S = \sum_{i=1}^{N-1} Z_i Z_i$$

$$(N-1) DSD' = D \left(\sum_{i=1}^{N-1} Z_i Z_i\right) D'$$

$$= \sum_{i=1}^{N-1} (DZ_i)(DZ_i)'$$

$$(N-1) S^* = \sum_{i=1}^{N-1} Z_1^* Z_i^*$$

$$Z_i^* = DZ_i \quad ; \quad Z_i \sim N(0, D\Sigma D')$$

$$\Rightarrow Z_i^* \sim N(0, \Sigma)$$

Consider the equality,

$$\mathcal{V}' \Sigma^{-1} \mathcal{V} = (D^{-1} \mathcal{V}^*)$$

$$= \mathcal{V}^{*\prime} (D \Sigma D')^{-1} \mathcal{L}^*$$

$$= \mathcal{L}^{*\prime} \mathcal{L}^*$$

We shall consider the orthogonal matrix Q of order $p \ni$ ith element of the first row of q.

$$q_{i1} = \frac{y_i^*}{\sqrt{y^{*\prime} y^*}}$$

$$y_i^* = q_{1i} \sqrt{y^* y^*}$$

(i.e) $q_{11}^2 + q_{12}^2 + \cdots + q_{np}^2 = 1$

$$\frac{y_1^{*2}}{y^* y^{*\prime}} + \cdots + \frac{y_p^{*2}}{y^* y^{*\prime}} = 1$$

$$\sum_{i=1}^{p} y_i^{*2} = 1$$

$$\frac{y^{*\prime} y^*}{y^{*\prime} y^*} = 1$$

The other $(p-1)$ rows can be defined by some arbitary rule.

Since q depends on y^* it is a matrix of a random matrix,

Let us define $U = q\, y^*$

$$\begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_p \end{bmatrix} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1p} \\ q_{21} & q_{22} & \cdots & q_{2p} \\ \vdots & & & \\ q_{p1} & q_{p2} & \cdots & q_{pp} \end{bmatrix} \begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_p^* \end{bmatrix}$$

$$U_1 = q_{11} y_1^* + q_{12} y_2^* + \cdots + q_{1p} y_p^*$$

$$U_2 = q_{21} y_1^* + q_{22} y_2^* + \cdots + q_{2p} y_p^*$$

$$\vdots$$

$$U_p = q_{p1} y_1^* + q_{2p} y_2^* + \cdots + q_{pp} y_p^*$$

$$y_i^* = q_{1i}\sqrt{y^{*\prime}y^*}$$

$$U_1 = \sqrt{y^{*\prime}y^*}\sum_{i=1}^{p} q_i^2$$

$$U_1 = \sqrt{y^{*\prime}y^*}$$

$$U_2 = q_{21}q_{11}\sqrt{y^{*\prime}y^*} + q_{22}q_{12}\sqrt{y^{*\prime}y^*} + \dots +$$

$$q_{2p}q_{1p}\sqrt{y^{*\prime}y^*}$$

$$= \sqrt{y^{*\prime}y^*}\sum_{i=1}^{p} q_{21}q_{1i} = 0$$

$$T^2 = \sqrt{y^{*\prime}S^{*-1}y^*}$$

$$T^2 = y^{*\prime}S^{*-1}y^*$$

$$\frac{T^2}{N-1} = \frac{1}{N-1}y^*S^{*-1}y^*$$

$$= y^{*\prime}[(N-1)S^*]^{-1}y^* \quad \begin{bmatrix} U = Qy^* \\ y^* = UQ^{-1} \end{bmatrix}$$

$$= U'(Q^{-1})'[(N-1)S^*]^{-1}Q^{-1}U$$

$$\frac{T^2}{N-1} = U'B'U$$

$$\left[\therefore B = Q(N-1)S^{*-1}Q\right]$$

$$= (U_1 0 \dots 0)\begin{bmatrix} b_{11} & b^{12} & \dots & b^{1p} \\ b^{21} & b^{22} & \dots & b^{2p} \\ \vdots & & & \\ b^{p1} & b^{p2} & \dots & b^{pp} \end{bmatrix} = \begin{bmatrix} U_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\frac{T^2}{N-1} = U_1^2 b^{11}$$

$$= \frac{U_1^2}{(b^{11})^{-1}}$$

$$\frac{T^2}{N-1} = \frac{y^{*\prime}y^*}{(b^{11})^{-1}}$$

$$Y^* \sim N(\mu^*, \Sigma) \Rightarrow U_1^2 \sim N^2(P-1)+1 = \chi'^2 P.$$

Appliation of Hotelling T^2 Statistics :

(i) T^2- Statistics is used to test hypothesis that the mean of one normal population equal to the mean of the other normal population of there where the co- variance matrices of there two population of assumes to the equal and unknown.

(ii) T^2- Statistics is used to test equality of the mean of Q- Samples.

(iii) T^2- Statistics is used to test two samples Problem is unequal Covariance matrix.

[T^2- Statistics is used to test to hypothesis of single mean vector and two mean vector with know of unknown Σ matrix].

Properties of T^2- Statistics :

T^2 is not only affected by changing origin of rresponse variate $X_1, X_2 \cdots X_n$ but the invariant under linear transformation. Where, $X \sim NP(\mu, \Sigma)$ if (c is non-Singular of $P \times P$ matrix) co-effient of λ is $P \times 1$ column the T^2 Statistics is $\bar{Y} = (C\bar{X} + \lambda)$.

$$T^2 = N[\bar{y} - E(\bar{y})]' S^{-1}{}_{Y} [\bar{y} - E(\bar{y})]$$

Substitute ; $\bar{Y} = (\bar{X} + \lambda)$

$$= N[(\bar{X}+\lambda) - (\mu+\lambda)' (SC')^{-1} (\bar{x}+\lambda) - (\mu+\lambda)]$$

$$= N[c'(\bar{x}-\mu)'] c^{-1} s^{-1} c' [(\bar{x}-\mu)]$$

$$= N (\bar{x}-\mu)' c' c^{-1} s^{-1} (\bar{x}-\mu)$$

$$T^2 = N (\bar{x}-\mu)' s^{-1} (\bar{x}-\mu)$$

T^2 is based on X.

Wischart Distribution :

In the Univariable case if $X_1, X_2 \cdots X_n$ are 'N' independent Observation then,

$$\hat{\sigma}^2 = s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})'$$

If we replace the Scalar by vector we get, the Multivariate analouge.

$$\hat{\Sigma} = S = \frac{1}{N-1} \Sigma (x_i - \bar{x})(x_i - \bar{x})'$$

$$\hat{\Sigma} = \frac{A}{N-1}$$

If the normal assume then, the distribution of A is nothing but the distribution of Sample co-variance Matrix is called Wishart distribution.

Properties of Wishart distribution:

(i) If A_1 is distributed as $W_{m_1}(A_1/\varepsilon)$ independently of A_2, which is distributed $W_{m_2}(A_2/\varepsilon)$ then, A_1+A_2 is distributed as Wishart with $W_{m_1+m_2}(A_1+A_2/\varepsilon)$.

(ii) If A is distributed as $W_m(A/\varepsilon)$ then, CAC' is distributed as $W_m(CAC'/C\varepsilon C')$.

Additive Property of Wishart distribution:

If A_i $(i=1,2 \ldots q)$ are independent distributed of $W(\Sigma, n_i)$ respectively then,

$$A = \sum_{i=1}^{q} A_i \sim W\left(\Sigma, \sum_{i=1}^{q} n_i\right)$$

Proof: We know that, $\phi_A(\theta) = \dfrac{|\Sigma^{-1}|^{n/2}}{|\Sigma^{-1} - 2i\theta|^{n/2}}$

$A_1 \sim W(\Sigma, n_1)$ then, $\phi_A(\theta) = \dfrac{|\Sigma^{-1}|^{n/2}}{|\Sigma^{-1} - 2i\theta|^{n/2}}$

Since A_1, A_2 are independent,

$$\phi_{A_1+A_2}(\theta) = \phi_{A_1}(\theta) + \phi_{A_2}(\theta)$$

$$= \frac{|\Sigma^{-1}|^{n_1/2}}{|\Sigma^{-1} - 2i\theta|^{n_1/2}} + \frac{|\Sigma^{-1}|^{n_2/2}}{|\Sigma^{-1} - 2i\theta|^{n_2/2}}$$

$$\phi_{A_1+A_2}(\theta) = \frac{|\Sigma^{-1}|^{\frac{n_1+n_2}{2}}}{|\Sigma^{-1} - 2i\theta|^{\frac{n_1+n_2}{2}}}.$$