PROBABILITY AND STATISTICS

PROBABILITY AND STATISTICS

Syllabus

Unit-I

Vector Spaces- Vector Spaces and subspaces-Null Spaces, Column Spaces and Linear Transformations. Linearly Independent sets- Bases- Coordinates systems.

Vector Spaces

In this chapter, vector space and a subspace of a vector space are defined, Null space, Column space in a vector space, Linear span, linear transformation in a vector space, and the bases of a vector space are also defined and described.

Unit-II

Probability- Basic terminology, Three types of Probability rules, Statistical independence, Statistical dependency, Bayes theorem.

Probability distributions - random variables, expected values, Binomial distribution, Poisson distribution, Normal distribution, choosing correct distribution.

Unit-III

Sampling and sampling distributions – Random sampling, Non-Random sampling distributions, operational considerations in sampling.

Estimation – Point Estimates, Interval Estimates, Confidence intervals, calculating interval estimates of the mean and proportions, t-distribution, determinates of sample size in estimation.

Unit-IV

Testing Hypothesis - one sample tests – Hypothesis testing of mean when the population standard deviation is known, powers of hypothesis test, hypothesis testing of proportions, Hypothesis testing of means when standard deviation is not known.

Testing Hypothesis – Two sample tests - Tests for difference between means - large sample, small sample, with dependent samples, testing for difference between proportions, Large sample.

Unit-V

Chi-square and Analysis of Variance – chi-square as test of independence, chi-square as a test of goodness of fit, analysis of variance, Inferences about a population variances.

Regression and Correlation – Simple Regression – Estimation using regression line, correlation analysis, making inferences about the population parameters, limitations, errors and caveats in regression and correlation analysis, multiple regression and correlation analysis, Finding multiple regression equations and making inferences about population parameters.

COURSE WRITERS

1. Dr. RAGHAVENDER SHARMA MAMILLAPALLY, M.Sc, Ph D.
Assistant Professor
Osmania University
Hyderabad-07

2. Dr. KAKULAPATI MURALI KRISHNA, M Sc, M Phil, Ph D.
Principal
G. Pulla Reddy Degree & P.G. College
Mehdipatnam, Hyderabad-28

3. Dr. M. JAGAN MOHAN RAO, M.Sc, Ph D.
Principal
St. Anthony's Degree College
Balapur, Hyderabad-58

4. Dr. M. UDAY SANKAR, M.Sc, Ph D.
Department of Statistics,
Osmania University
Hyderabad-07

UNIT-I

Vector Spaces- Vector Spaces and subspaces-Null Spaces, Column Spaces and Linear Transformations. Linearly Independent sets- Bases- Coordinates systems.

Vector Spaces

In this chapter, vector space and a subspace of a vector space are defined, Null space, Column space in a vector space, Linear span, linear transformation in a vector space, and the bases of a vector space are also defined and described

Scalar : A physical quantity which has only magnitude is called a scalar.
Examples: Length, volume, time.

Vector: A vector is a physical quantity which has both magnitude and direction.
Geometrically a directed line segment is called a vector.
Examples: Force, Velocity, acceleration.

Note: All real numbers are scalars.

Vector Spaces

A Vector space is a non-empty set V of objects called vectors on which are defined two operations called addition and multiplication by scalars (real numbers) subject to the axioms (or rules) listed below
These axioms must hold for all vectors $\bar{u}, \bar{v}, \bar{w}$ in V and for all scalar c and d.

1) The sum of u and v denoted by $\bar{u}+\bar{v}$ is in v i.e $\bar{u},\bar{v} \in \bar{v}$(Closer property)
2) $\bar{u} + \bar{v}=\bar{v} + \bar{u}$(commutative or abelian property)
3) $(\bar{u}+\bar{v})+\bar{w}=\bar{u}+(\bar{v}+\bar{w})$(Associative property)
4) There is a zero vector 0 in V such that $\bar{u}+0=0+\bar{u}=\bar{u}$(identity element)
5) For each $\bar{u}$ in V there is a vector $-\bar{u} \in V \ni \bar{u} + (-\bar{u}) = 0$ (inverse element)

6) The scalar multiplication

$$c \in F, \bar{u} \in \bar{v}$$

$$\Longrightarrow c\bar{u} \in \bar{v}$$

7) c($\bar{u}$+v)=c$\bar{u}$+c$\bar{v}$ (u,v are vectors) $c \in F, \bar{u} \in \bar{v}$ (c is scalar)

8) (c+d)u=c$\bar{u}$+d$\bar{u}$ (u is vectorc, d $\in F$, $\bar{u} \in \bar{v}$

9) c(d$\bar{u}$) =(cd) $\bar{u}$, $c, d \in F, \bar{u} \in \bar{v}$

10) $1\bar{u}=\bar{u}$

Problems

1Q) Show that R^3 with usual notation addition and scalar multiplication is a vector space over R

Sol: R^3= {(a, b, c) : a, b, c $\in R$}

Here scalar are real numbers

Let u, v, w $\in R^3$

And $\bar{u}= (x_1, x_2, x_3)$ $x_1, x_2, x_3 \in R$

$\bar{v}=(y_1, y_2, y_3)$ $y_1, y_2, y_3 \in R$

$\bar{w}=(z_1, z, z_3) z_1, z, z_3 \in R$

(i) The sum of u and v denoted by $u + v$ is in v ieu, $v \in V \Rightarrow u + v \in v$ (closure property)

$\bar{u}=(x_1, x_2, x_3)$

$\bar{v}=(y_1, y_2, y_3)$

$\bar{u} + \bar{v}= (x_1, x_2, x_3) +(y_1, y_2, y_3)$

$=(x_1 + y_1, x_2 + y_2, x_3 + y_3)$

$x_1, x_2 \in R \Rightarrow x_1 + x_2 \in R$

$y_1, y_2 \in R \Rightarrow y_1 + y_2 \in R$

$z_1, z_2 \in R \Rightarrow z_1 + z_2 \in R$

$x_1, y_1 \in R \Rightarrow x_1 + y_1 \in R$

$x_2, y_2 \in R \Rightarrow x_2 + y_2 \in R$

$x_3, y_3 \in R \Rightarrow x_3 + y_3 \in R$

$\therefore \bar{u} + \bar{v}$

$=(x_1 + y_1, x_2 + y_2 + x_3 + y_3) \in R^3$

(ii) $u + v = v + u$ (Commutative or abelian property)

$\bar{u}, \bar{v} \in V \Rightarrow \bar{u} + \bar{v} = \bar{v} + \bar{u}$

$\bar{u}=(x_1, x_2, x_3)$

$\bar{v}=(y_1, y_2, y_3)$

$\bar{u} + \bar{v}= (x_1, x_2, x_3) +(y_1, y_2, y_3)$

$=(x_1 + y_1, x_2 + y_2 + x_3 + y_3)$

In R $x_1 + y_1 = y_1 + x_1$

$x_2 + y_2 = y_2 + x_2$

$x_3 + y_3 = y_3 + x_3$

$= (y_1 + x_1, y_2 + x_2, y_3 + x_3)$

$= (y_1, y_2, y_3) +(x_1, x_2, x_3)$

$= \bar{v} + \bar{u}$

$\therefore\ \bar{u} + \bar{v} = \bar{v} + \bar{u}$

(iii) $(\bar{u} + \bar{v}) + \bar{w} = \bar{u} + (\bar{v} + \bar{w})$ (Associative Property)

$(\bar{u} + \bar{v}) + \bar{w} = [(x_1, x_2, x_3) +(y_1, y_2, y_3)] + (z_1, z_2, z_3)$

$= (x_1 + y_1, x_2 + y_2 + x_3 + y_3) + (z_1, z_2, z_3)$

$= [(x_1 + y_1) + z_1, (x_2 + y_2) + z_2, (x_3 + y_3) + z_3]$

$x_1, y_1, z_1 \in R$

$\Rightarrow (x_1 + y_1) + z_1 = x_1 + (y_1 + z_1)$

$\Rightarrow (x_2 + y_2) + z_2 = x_2 + (y_2 + z_2)$

$\Rightarrow (x_3 + y_3) + z_3 = x_3 + (y_3 + z_3)$

$\Rightarrow x_1 + (y_1 + z_1), x_2 + (y_2 + z_2), x_3 + (y_3 + z_3)$

$\Rightarrow (x_1, x_2, x_3) + [(y_1 + z_1), (y_2 + z_2), (y_3 + z_3)]$

$\Rightarrow [(x_1, x_2, x_3) +(y_1, y_2, y_3)] + (z_1, z_2, z_3)$

$\Longrightarrow \bar{u} + (\bar{v} + \bar{w})$

(iv) Identity element: $0+\bar{u} = \bar{u} + 0 = \bar{u}$ (Zero vector $\in V$

We have $0 \in R$

$\Longrightarrow (0, 0, 0) \in R^3$

Let $\bar{0} = (0, 0, 0)$

Consider

$$\bar{0} + \bar{u} = (0, 0, 0) + (x_1, x_2, x_3)$$
$$= (0 + x_1, 0 + x_2, 0 + x_3)$$
$$= (x_1, x_2, x_3) = \bar{u}$$

(v) For each u in v there is a vector- $u \in v \ \exists \ u + (-u) = 0$ (inverse element)

We have $\bar{u} = (x_1, x_2, x_3)$

$x_1, x_2, x_3 \in R$

$\Longrightarrow$ - $x_1, -x_2, -x_3 \in R$

Define $\bar{v} = (-x_1, -x_2, -x_3)$

$\bar{u} + \bar{v} = (x_1, x_2, x_3) + (-x_1, -x_2, -x_3)$

$= (x_1 - x_1, x_2 - x_2, x_3 - x_3)$

$= (0, 0, 0) = \bar{0}$

$\therefore \bar{v}$ is inverse of $\bar{u}$

vi) The scalar multiplication

Define Scalar multiplication

$c\bar{u} = (x_1, x_2, x_3)$

$= (cx_1, cx_2, cx_3) \in V \ (cx_1, cx_2, cx_3 \ are \ in \ R \ )$

The scalar multiplication

$c \in R \ , u \in v$

$c\bar{u} \in v$

$\bar{u} = (x_1, x_2, x_3)$

Where $x_1,\ x_2,\ x_3 \in R$

Here c is a scalar and $c \in R$

Consider $c\bar{u} = c(x_1, x_2, x_3)$

$= (cx_1, cx_2, cx_3)$

Here $cx_1,\ cx_2,\ cx_3 \in R$

$\therefore c\bar{u} \in R^3$

vii) C $(\bar{u} + \bar{v}) = C\bar{u} + C\bar{v}$

Let $\bar{u} = (x_1, x_2, x_3)$

$\bar{v} = (y_1, y_2, y_3)$

Consider

$$C(\bar{u} + \bar{v} = C((x_1, x_2, x_3) + (y_1, y_2, y_3))$$
$$= C(x_1 + y_1, x_2 + y_2, x_3 + y_3)$$
$$= (C(x_1 + y_1), C(x_2 + y_2), C(x_3 + y_3))$$
$$= (Cx_1 + Cy_1, Cx_2 + Cy_2, Cx_3 + Cy_3)$$
$$= (Cx_1, Cx_2, Cx_3) + (Cy_1, Cy_2, +Cy_3)$$
$$= C(x_1, x_2, x_3) + C(y_1, y_2, y_3)$$
$$= C\bar{u} + C\bar{v}$$

viii) (c+d) $\bar{u} = c\bar{u} + d\bar{u} \Rightarrow c, d \in R, \bar{u} \in R^3$

Let $\bar{u} = (x_1, x_2, x_3)$

Consider

$$= (c + d)\bar{u} = (c + d)(x_1, x_2, x_3)$$
$$= ((c + d)x_1, (c + d)x_2, (c + d)x_3)$$
$$= (cx_1 + dx_1), (cx_2 + dx_2), (cx_3 + dx_3)$$
$$= (\bar{c}x_1, \bar{c}x_2, cx_3) + (dx_1, dx_2, dx_3)$$
$$= c(x_1, x_2, x_3) + d(x_1, x_2, x_3)$$
$$= c\bar{u} + d\bar{u}$$

ix) $c(d\bar{u}) = (cd)\overline{u}$

Here c, d, $\in R$

$\bar{u} = (x_1, x_2, x_3) \in R^3$

Consider

$c(d\bar{u}) = (cd(x_1, x_2, x_3))$

$= c(\mathrm{d}x_1, dx_2, dx_3)$

$= c(\mathrm{d}x_1), c(dx_2), c(dx_3)$

$= c, d, x, \in R \;\; c(\mathrm{d}x_1) = (cd)x_1$

$= ((cd)x_1, (cd)x_2, (cd)x_3)$

$= cd(x_1, x_2, x_3)$

$= \mathrm{cd}\ (\bar{u})$

$\therefore c(d\bar{u}) = (cd)\bar{u})$

x) $\bar{u} = \bar{u}$

We have 1 , $\in R$

$\bar{u}$ =(x_1, x_2, x_3)

Here $1x_1, 1x_2, 1x_3 \in R$

Also $1x_1 = x_1$

$1x_2 = x_2$

$1x_3 = x_3$

$\bar{u} = 1(x_1, x_2, x_3)$

$= (1x_1, 1x_2, 1x_3) = (x_1, x_2, x_3) = \bar{u}$

$\therefore R^3(R)$ *is a vector space.*

2Q) $\bar{u} + \bar{v} = \bar{v} + \bar{u}$ **(commutative or Abelian property)**

Sol: $A + B = B + A$

$$A = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix} \qquad B = \begin{bmatrix} c_1 & c_2 & c_3 \\ d_1 & d_2 & d_3 \end{bmatrix}$$

$$A + B = \begin{bmatrix} a_1 + c_1 & a_2 + c_2 & a_3 + c_3 \\ b_1 + d_1 & b_2 + d_2 & b_3 + d_3 \end{bmatrix}$$

In R, since

$$\begin{aligned} a_1 + c_1 &= c_1 + a_1 \\ a_2 + c_2 &= c_2 + a_2 \\ a_3 + c_3 &= c_3 + a_3 \end{aligned}$$

$$\begin{aligned} b_1 + d_1 &= d_1 + b_1 \\ b_2 + d_2 &= d_2 + b_2 \\ b_3 + d_2 &= d_3 + d_3 \end{aligned}$$

$$= \begin{bmatrix} c_1 + a_1 & c_2 + a_2 & c_3 + a_3 \\ d_1 + b_1 & d_2 + b_2 & d_3 + d_3 \end{bmatrix}$$

$$= \begin{bmatrix} c_1 & c_2 & c_3 \\ d_1 & d_2 & d_3 \end{bmatrix} + \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix}$$

$$= \text{B+A}$$

$$\therefore A + B = B + A$$

We know that matrix addition is associative

(i) $(u + v) + w = u + (v + w)$

We know that matrix addition is associative

$\therefore A + (B + C) = (A + B) + C$

(ii) Identify $0 + \bar{u} = \bar{u} + 0 = \bar{u}$

A+ 0=A

We have $0 \in R$

Define

$$0=\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

Consider A+0 $$=\begin{bmatrix}a_1 & a_2 & a_3\\ b_1 & b_2 & b_3\end{bmatrix}+\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$=\begin{bmatrix}a_1+0 & a_2+0 & a_3+0\\ b_1+0 & b_2+0 & b_3+0\end{bmatrix}=\begin{bmatrix}a_1 & a_2 & a_3\\ b_1 & b_2 & b_3\end{bmatrix}=A$$

$\therefore$ *is the identity element*

(iii) $x\,u \in R$

Let x u $\in R$

Consider

$$Xa = x\begin{bmatrix}a_1 & a_2 & a_3\\ b_1 & b_2 & b_3\end{bmatrix}$$

$$= \begin{bmatrix}xa_1 & xa_2 & xa_3\\ xb_1 & xb_2 & xb_3\end{bmatrix}$$

Here $\{ x, a_1 \in R \Rightarrow xa_1 \in R$

$x, a_2 \in R \Rightarrow xa_2 \in R$

$x, a_3 \in R \Rightarrow xa_3 \in R$

$x, b_1 \in R \Rightarrow xb_1 \in R$

$x, b_2 \in R \Rightarrow xb_2 \in R$

$x, b_3 \in R \Rightarrow xb_3 \in R\}$

$\therefore x\,A \in V$

(iv) $x\,(u+v) = xu + xv$

Consider

$$\mathrm{X}(A+B)$$

$$= x\begin{bmatrix}a_1+c_1 & a_2+c_2 & a_3+c_3\\ b_1+d_1 & b_2+d_2 & b_3+d_3\end{bmatrix}$$

$$= \begin{bmatrix} x(a_1 + c_1) & x(a_2 + c_2) & x(a_3 + c_3) \\ x(b_1 + d_1) & x(b_2 + d_2) & x(b_3 + d_3) \end{bmatrix}$$

$$= \begin{bmatrix} xa_1 + xc_1 & xa_2 + xc_2 & xa_3 + xc_3 \\ xb_1 + xd_1 & xb_2 + xd_2 & xb_3 + xd_3 \end{bmatrix}$$

$$= \begin{bmatrix} xa_1 & xa_2 & xa_3 \\ xb_1 & xb_2 & xb_3 \end{bmatrix} + \begin{bmatrix} xc_1 & xc_2 & xc_3 \\ xd_1 & xd_2 & xd_3 \end{bmatrix}$$

$$= Xa + Xb$$

3Q) Show that the set H of all points of R^2 of the form $(3s, 2 + 5s)$ is not a vector space by showing that it is not closed under scalar multiplication

Sol: $V = \{(3s, 2 + 5s) : S \epsilon R\} = H$

$\bar{u} = (35, 2 + 5s)$

For $s = 1, \bar{u} = (3, 2 + 5)$

$= (3,7)$

$Let\ \ c = 3,\ c\bar{u} = 3(3,7)$

$= (9, 21)$

$If\ 3\bar{u} \in H$

$The\ 3s, 2 + 5s\ ) = (9,21)$

$\Rightarrow 3s = 9\ , 2 + 5s = 21$

$\Rightarrow s = 3 \quad 5s = 21 - 2 = 19$

$s = \frac{19}{5}$

$This\ is\ not\ possible$

$Therefore \quad H\ is\ not\ closed\ under\ scalar\ multiplication. The\ problem$

4Q) ***Let H*** **be the set of points inside and on the unit circle in** $xy - plane$ **that is** ***Let H*** $= \{(x, y): x2 + y2 \leq 1\}$. **Find specific examples of two vectors and a scalar to show that H is not a vector space.**

Sol: Given $\bar{u}$ = $(0,1),\ 0^2 + 1^2 = 1 \leq 1$

$\bar{v}$ = $(1,0), 1^2 + 0^2 = 1 \leq 1$ $\quad \bar{u}, \bar{v} \in H$

$\bar{u} + \bar{v}$ = $(0,1) + (1,0)$

= $(1,1)$

Here $1^2 + 1^2 = 2 \leq 1$ is wrong

$$\therefore \bar{u} + \bar{v} \notin H$$

H is not closed under addition.

Scalar multiplication

$H = \{(X, Y)\ x2 + y2 \leq 1\}$

$Let\ \bar{u} = (0,1)$

$C = 2$

$2\bar{u} = 2(0, 1) = (0, 2)$

$0^2 + 2^2 = 4 \leq 1$is wrong

$c\ \bar{u} \notin H$

$\therefore\ Scalar\ multiplication\ \ fails$

$\therefore\ H\ \ is\ not\ a\ vector\ space$

<u>Subspaces:</u>

A Subspace of a vector space V is a subset H of V that has the three following properties

(i) $The\ zero\ vector\ of\ V\ is\ in\ H\ i.e\ \bar{0} \in H$

(ii) $If\ \bar{u}.\bar{v}\ \in H\ then\ \bar{u} + \bar{v}\ \in H$

(iii) $\bar{u} \in Hand\ c\ is\ any\ scalar\ then\ c\ \bar{u} \in H$

Note:

Let $\bar{u}, \bar{v}, \bar{w} \in H$

$\Rightarrow \bar{u}, \bar{v}, \bar{w} \in V$ $(H \sqsubseteq V)$

$(\bar{u} + \bar{v}) + \bar{w} = \bar{u} + (\bar{v} + \bar{w})$

And $\bar{u} + \bar{v} = \bar{v} + \bar{u}$

We have $\bar{0} \in H$

$c\,\bar{u} \in H$

$Let\ c = -1$

$(-1)\,\bar{u} = -\bar{u} \in H$

$H\ is\ subset\ of\ V$

$\therefore \bar{u} \in H \Rightarrow \bar{u} \in V$

$\therefore C(\bar{u} + \bar{v}) = C\bar{u} + C\bar{v}$

$(c + d)\bar{u} = C\bar{u} + d\bar{u}$

$(cd)\,\bar{u} = c(d\bar{u})$

$We\ have\ C\bar{u}\ H$

$Put\ c = 1$

$\Rightarrow 1\bar{u} = \bar{u} \in H$

Theorem: A Subset W of a vector space V is a subspace of V if and only if the following condition holds.

If $\bar{u}\ and\ \bar{v}$ are any 2 vectors in w then for any 2 scalar c, d the vector $c\bar{u} + d\bar{u} \in$ W

Proof: Given V is a vector space W is subset of V.

Suppose W is subspace of V

(1) $\bar{0} \in W$

(2) $\bar{u}, \bar{v} \in W \Rightarrow \bar{u} + \bar{v} \in W$

(3) $c\bar{u} \in$W

$Let\ c, d\ be\ any\ two\ scalars$

$Let\ \bar{u}, \bar{v} \in W$

$From\ 3)\ \therefore\Rightarrow c\bar{u}, d\bar{v}\ \in W$

$c\bar{u}, d\bar{v}\ \in$W

$From\ 2)\ \Rightarrow c\bar{u}, d\bar{v}\ \in W$

$Conversely\ suppose\ that$

$c\bar{u} + d\bar{v}\ \in wwhere\ c, d\ are\ scalars$

$c\bar{u} + d\bar{v}\ \in W \rightarrow 1$

$Put\ c = 0, d = 0\ in\ 1$

$\Rightarrow o\bar{u} + o\bar{v}\ \in W$

$\Rightarrow \bar{o} \in W$

$Put\ c = 1, d = 1\ in\ 1$

$1\bar{u} + 1\bar{v}\ \in W$

$\bar{u} + \bar{v}\ \in W$

$Put\ c = c\ , d = 0\ in\ 1$

$c\bar{u} + o\bar{v}\ \in W$

$\Rightarrow c\bar{u} \in W$

$Therefore\ W\ is\ subspace\ of\ V$

$Hence\ proved$

Problems:

Q) Consider the vector space R^2 with usual addition and scalar multiplication.

$R^2=\left\{\begin{bmatrix} x \\ y \end{bmatrix} : x, y\ \in R\right\}$ S. T the subset

$H=\left\{\begin{bmatrix} x \\ -x \end{bmatrix} : x, \in R\right\}\ is\ a\ subspace\ of\ R^2$

Sol: Given $R^2 = \left\{\begin{bmatrix} x \\ y \end{bmatrix} : x, y\ \in R\right\}$

$H=\left\{\begin{bmatrix} x \\ -x \end{bmatrix} : x, \in R\right\}$ To show that

H is a subspace of R^2.

(1) We have $O \in R$

$$\Rightarrow \bar{0} = \begin{bmatrix} 0 \\ -0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \in H$$

(2) Let $\bar{u} = \begin{bmatrix} x \\ -x \end{bmatrix} \in H$

$$\bar{v} = \begin{bmatrix} y \\ -y \end{bmatrix} \in H$$

Consider

$$\bar{u} + \bar{v} = \begin{bmatrix} x \\ -x \end{bmatrix} + \begin{bmatrix} y \\ -y \end{bmatrix}$$

$$= \begin{bmatrix} x + y \\ -x - y \end{bmatrix}$$

$$= \begin{bmatrix} x + y \\ -x - y \end{bmatrix} \in H$$

$$x, y \in R$$

$$\Rightarrow x + y \in R \ \& \ (-x + y) \in R]$$

(3) *Let* $\bar{u} = \begin{bmatrix} x \\ -x \end{bmatrix}$

Let c be any scalar

$$c\bar{u} = c \begin{bmatrix} x \\ -x \end{bmatrix} = \begin{bmatrix} cx \\ -cx \end{bmatrix} \in H$$

$$[c_1 x \in R \Rightarrow c\, x \in R, -cx \in R]$$

Therefore H is a subspace of R^2

I. Theorem:

The intersection H∩k of two subspaces H and K of V is a subspace of V, But the union H∪K need not necessarily be a subspace of V.

Proof:

Given v is a vector space, H, K *are two subspaces of* V.

$\therefore \bar{o} \in H$

$\forall \bar{u}, \bar{v} \in H, \bar{u} + \bar{v} \in H$

$and\ for\ c\ any\ scalar, \bar{u} \in H \Longrightarrow c\bar{u} \in H$

$Similarly\ K\ is\ a\ subspace\ of\ V.$

→ $\bar{o} \in K$

$\bar{u} + \bar{v} \in K, \bar{u}, \bar{v} \in K$

$c\bar{u} \in K$c is any scalar $\bar{u} \in K$

$To\ Prove\ that\ H \cap k\ is\ subspace\ of\ V$

$we\ have\ \bar{o} \in H, \bar{o} \in K \Longrightarrow \bar{o} \in H \cap k$ – (1)

$Let\ \bar{u}, \bar{v} \in H \cap k$

→ $\bar{u}, \bar{v} \in H \cap k\ and\ \bar{u}, \bar{v} \in k$

→ $\bar{u} + \bar{v}$ ∈H∩k$\bar{u} + \bar{v}$ ∈k

→ $\bar{u} + \bar{v}$ ∈H∩k – (2)

Let c be any scalar and $\bar{u}$ ∈H∩k

→ $\bar{u} \in H\ and\ \bar{u} \in K$

$\therefore c\bar{u}$ ∈ Hand $c\bar{u}$ ∈ K

→ $c\bar{u}$ ∈ H∩k – (3)

$Therefore\ from$ 1,2 and 3 $H \cap k\ is\ a\ subspace\ of\ V.$

To prove that $H \cup K$ need not necessarily be a subspace of V it is enough to give one example

$Let\ V = R2$

$We\ know\ that\ H = \{(x1 - x), x \in R\}\ and\ K = \{(x, 2x) : x \in R\}\ are\ subspaces\ of\ V$

$H \cup K = \{(x1 - x)(x, 2x), x \in R\}$

$(2_1 - 2) \in H \cup K$

$(2_1\ 4) \in H \cup K$

$But\ (2,-2)\ +\ (2,4)\ =\ (2\ +\ 2,-2\ +\ 4)\ =\ (4,2)\ \notin H \cup K$

$\therefore$ Here $(4,2)$ is not of the form $(x,-x)\ or\ (x,2x)$

$\therefore$ Vector addition fails

$\therefore$ $H \cup K$ need not necessarily be a subspace of V

Hence the Theorem.

NOTE:

$H \cup K$ is subspace of V if and only if $H \sqsubseteq K$ or $K \sqsubseteq H$

1) H, K are Subspaces $H \sqsubseteq K$

Definition:

Let H and K be two subspaces of a vector space V then the Sum of H and K written as H + K is defined as.

$H + K\ =\ \{w: w = u + v;\ u \in H, v \in K\}$

$H + K\ =\ \{u + v;\ u \in H, v \in K\}$

Theorem: Given Subspaces H and K of a vectors of $a\ vectorspace\ v$, the sum H and K written as $H + K$ is a subspace of V.

Proof:

Given V is a vector space. H is subspace of V

K is subspace of V we have $H + K\ =\ \{w;\ w = u + v;\ u \in H, v \in K\}$

Here H and K are subset of V –

→ $H + K$ is subsets of V

H is subspace of V

→ 1) $\bar{o} \in H$

2) $\bar{u} + \bar{v} \in H, \bar{u}, \bar{v} \in H$

3) $c\bar{u} \in H, \bar{u} \in H, c$ is any scalar

K is subspace of V

→ 1) $\bar{o} \in K$

2) $\bar{u} + \bar{v} \in K, \bar{u}, \bar{v} \in K$

3) $c\bar{u} \in K, \bar{u} \in K, c$ is any scalar

To prove that H+K is a subspace of V

We have $\bar{o} = \bar{o} + \bar{o} \in H + K$

→ $\bar{o} \in H + K$ – (1)

Let $w_1, w_2 \in H + K$

→ $w_1 = u_1 + v_1 , u_1 \in H, v_1 \in H, v, \in K$

$w_2 = u_2 + v_2 \ \in H + K$

$u_2 \in H, v_2 \ \in H$

Consider

$w_1 + w_2$ = $(u_1 + v_1) + (u_2 + v_2)$

= $(u_1 + \ u_2) + (v_1 + v_2)$

$u_1, u_2 \in H, H \ is \ subspace$

$\Rightarrow u_1 + \ u_2 \in H$

= $v_1, v_2, \in K, K$ is subspace

$\Rightarrow v_1 + \ v_2 \in K$

$\therefore w_1, w_2 \in H + K \longrightarrow 2$

Consider

$c\ w_1 = c(u_1 + v_1) = cu_1 + cv_1$

$cu_1 \in H, cv_1 \in K$

$\therefore cw_1 \in H + K \longrightarrow 3$

From 1, 2 & 3 $H + K$ is subspace of V

Note: $H = \{w = u + o; u \in H \}$ $O \in K$

$\Rightarrow H \subseteq H + K$

$K = \{w = 0 + V; O \in H, V \in K \}$

$\Rightarrow \ K \subseteq H + K$

Here H, K are subset of $H + K$ also H, K are subspaces of $H + K$

Linear Combinations and span

Linear Combinations:

Given vectors $\overline{v_1}, \overline{v_2} \ldots\ldots \bar{v}_n$ and given scalars $c_1, c_2 \ldots\ldots c_n$ the vector y defined by $\bar{y} = c_1 + \overline{v_1} + c_2 + \overline{v_2} + \cdots .. + c_n \bar{v}_n$ is called a linear combination (L.C) of $\overline{v_1}, \overline{v_2} \ldots\ldots \bar{v}_n$ with weights (coefficients) $c_1, c_2 \ldots\ldots c_n$

Example:

1) $(a, b) = a(1,0) + b(1,0)$

$$v\_1 = (1,0), v\_2\ (0,1)$$

$(a, b)\ is\ L.C\ of\ \{(1,0), (0,1)\}$

2) Spam

Let $\overline{v_1}, \overline{v_2} \ldots\ldots \bar{v}_n$ any set of n vectors then the set of all the linear combinations of $\overline{v_1}, \overline{v_2} \ldots\ldots \bar{v}_n$ denoted by spam $\{\overline{v_1}, \overline{v_2} \ldots\ldots \bar{v}_n\}$ and is called the set spammed by $\overline{v_1}, \overline{v_2} \ldots\ldots \bar{v}_n$

Note:

1) Spam $\{\overline{v_1}, \overline{v_2} \ldots\ldots \bar{v}_n\}$ is the collection of all vectors that can be written in the form $c_1 \overline{v_1} + c_2 + \overline{v_2} + \epsilon \ldots . + c_n \bar{v}_n$ where $c_1, c_2 \ldots\ldots c_n$ are scalars

2) Spam $\{\overline{v_1}, \overline{v_2} \ldots\ldots \bar{v}_n\}$ contains every scalar multiple of each vector take $\overline{v_1}$, we can write

$$\overline{v_1} = 1.\overline{v_1} + 0.\overline{v_2} + \cdots .. 0\overline{v_n}$$

3) $\bar{0} \epsilon$ span $\{\overline{v_1}, \overline{v_2} \ldots\ldots \bar{v}_n\}$

$\bar{0} = 0\overline{v_1} + 0\ \overline{v_2} + \cdots .. + 0\overline{v_n}$

1Q) **If** $\overline{v_1} = \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix}$ ***and*** $\overline{v_2} = \begin{bmatrix} -4 \\ 3 \\ 8 \end{bmatrix}, \overline{v_3} = \begin{bmatrix} 2 \\ 5 \\ -4 \end{bmatrix}$ **determine whether =** $\begin{bmatrix} 3 \\ -7 \\ -3 \end{bmatrix}$ **is a linear combination of** $\overline{v_1}, \overline{v_2}, \overline{v_3}$ **or not**

Sol: We have to find whether scalars x_1, x_2, x_3 exits such that

$$x_1\overline{v_1} + x_2\overline{v_2} + x_3\overline{v_3} = \begin{bmatrix} 3 \\ -7 \\ -3 \end{bmatrix}$$

$$i.e x_1 \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix} + x_2 \begin{bmatrix} -4 \\ 3 \\ 8 \end{bmatrix} + x_3 \begin{bmatrix} 2 \\ 5 \\ -4 \end{bmatrix} = \begin{bmatrix} 3 \\ -7 \\ -3 \end{bmatrix}$$

$$= \begin{bmatrix} x_1 & -4x_2 & 2x_3 \\ 0x_1 & 3x_2 & 5x_3 \\ -2x_1 & 8x_2 & -4x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ -7 \\ -3 \end{bmatrix}$$

$$Let\ [A/B] = \begin{bmatrix} 1 & -4 & 2 & 3 \\ 0 & 3 & 5 & -7 \\ -2 & 8 & -4 & -3 \end{bmatrix}$$

$\bar{b}$ is linear combination of $\overline{v_1}, \overline{v_2}, \overline{v_3}$ if the system has a solution $R_3 \longrightarrow R_3 + 2R_1$

$\begin{bmatrix} 1 & -4 & 2 & 3 \\ 0 & 3 & 5 & -7 \\ 0 & 0 & 0 & -3 \end{bmatrix}$ The third row of the matrix indicates that the system is inconsistent

Hence $\bar{b}$ cannot be expressed as linear combination of $\overline{v_1}, \overline{v_2}, \overline{v_3}$.

Subspace Spanned of a Set

If $\overline{v_1}, \overline{v_2}, \overline{v_3}$……. v_n are any n vectors then the set of all vectors that can be written as Linear combination of these vectors is denoted by span $\{\overline{v_1}, \overline{v_2} \dots \dots \dots \overline{v_n}\}$

We call span $\{\overline{v_1}, \overline{v_2} \dots \dots \overline{v_n}\}$ as the set spanned by the set of vectors $\{\overline{v_1}, \overline{v_2} \dots \dots \overline{v_n}\}$

→ If it is a subspace say H of v then a spanning set for H is the set $\{\overline{v_1}, \overline{v_2} \dots \dots \overline{v_n}\}$ in H. Such that H= span $\{\overline{v_1}, \overline{v_2} \dots \dots \overline{v_n}\}$

2Q) Let w be the set of all vectors of the from $\begin{bmatrix} a-b \\ b-c \\ c-a \\ b \end{bmatrix}$ where a, b, c represent arbitrary real number find a *set S of vectors that span w*.

Sol: *Given*

$$\text{w} = \left\{ \begin{bmatrix} a-b \\ b-c \\ c-a \\ b \end{bmatrix}, a, b, c \in R \right\}$$

We have

$$\begin{bmatrix} a-b \\ b-c \\ c-a \\ b \end{bmatrix} = a\begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} + b\begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix} + c\begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} a-b \\ b-c \\ c-a \\ b \end{bmatrix} = a\overline{v_1} + \text{b}\overline{v_2} + c\overline{v_3}$$

$$\text{Span w} = \{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$$

3Q) Let w be the set of all vectors of the form $\begin{bmatrix} 3a+b \\ 4 \\ a-5b \end{bmatrix}$ where a, b, are arbitrary real number verity whether w form a vectors space.

Sol: Given

$$\text{w} = \left\{ \begin{bmatrix} 3a+b \\ 4 \\ a-5b \end{bmatrix}, a, b \in R \right\}$$

$$\begin{bmatrix} 3a+b \\ 4 \\ a-5b \end{bmatrix} = a\begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix} + b\begin{bmatrix} 1 \\ 0 \\ -5 \end{bmatrix} + 4\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$= = a\overline{v_1} + \text{b}\overline{v_2} + c\overline{v_3}$$

If zero vector is in w then $\begin{bmatrix} 3a+b \\ 4 \\ a-5b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

Here 4=0 is absurd therefore w is not a vector space.

4Q) ***Let*** $\overline{v_1} = \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \\ \mathbf{-1} \end{bmatrix}, \overline{v_2} = \begin{bmatrix} \mathbf{2} \\ \mathbf{1} \\ \mathbf{3} \end{bmatrix}, \overline{v_3} = \begin{bmatrix} \mathbf{4} \\ \mathbf{2} \\ \mathbf{6} \end{bmatrix}$ **and** $\bar{w} = \begin{bmatrix} \mathbf{3} \\ \mathbf{1} \\ \mathbf{2} \end{bmatrix}$

1) is $\bar{w}$ spanned $\{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$? How many vectors are in $\{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$?

2) How many vectors are in span $\{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$?

3) is $\bar{w}$ in the subspace? Why?

Sol: Given

$$\bar{w} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}, \overline{v_1} = \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix}, \overline{v_2} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}, \overline{v_3} = \begin{bmatrix} 4 \\ 2 \\ 6 \end{bmatrix}$$

Here $\bar{w} = \overline{v_1} + \overline{v_2}$

$\therefore \bar{w} = \text{span } \{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$

We got infinitely many linear combinations

5Q) ***Let w*** **be the set of all vectors at the form** $\begin{bmatrix} \mathbf{-a+1} \\ \mathbf{a-6b} \\ \mathbf{2b+a} \end{bmatrix}$ **where** $\boldsymbol{a, b, c \in R}$

find a set *S* of Vectors that span w or give an example to show that w is not a vector space.

Sol: Given

$$\begin{bmatrix} -a+1 \\ a-6b \\ 2b+a \end{bmatrix} = \begin{bmatrix} -a+0b+1 \\ a-6b \\ a+2b \end{bmatrix}$$

$$= a\begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} + b\begin{bmatrix} 0 \\ -6 \\ 2 \end{bmatrix} + 1\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$= a_1\overline{v_1} + \mathrm{b}\overline{v_2} + c_3\overline{v_3}$$

Let $H = \{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$

Here $\bar{0} \notin H$

$\therefore$ H is not a vector space

Theorem: ***Let $\overline{v_1}, \overline{v_2}$ be two vectors in a vector space V then H = Span $\{\overline{v_1}, \overline{v_2}\}$ is a subspace of V***

Proof:

Given V is a vector space.

$$H = \text{Span } \{\overline{v_1}, \overline{v_2}\}$$
$$= \{c_1\bar{v}_1 + c_2\bar{v}_2\} \therefore c_1, c_2 \in R$$

Now we have to prove that H is subspace of V

1) We have $\bar{o} = o\bar{v}_1 + o\overline{v_2}$

$\Longrightarrow \bar{o} \in R$

2) Let $\bar{u} = a\overline{v_1} + b\overline{v_2}$, $a, b \in R$

$\bar{v} = c\overline{v_1} + d\overline{v_2}$, $c, d \in R$

Consider $\bar{u} + \bar{v} = (a\overline{v_1} + b\overline{v_2}) + (c\overline{v_1} + d\overline{v_2})$

$= (a + c)\overline{v_1} + (b + d)\overline{v_2}$

Here $a, c \in R \Longrightarrow a + c \in R$

$b, d \in R \Longrightarrow b + d \in R$

$\therefore \bar{u} + \bar{v} \in H$ – (2)

3) *Let* $\bar{u} = a\overline{v_1} + b\bar{v}_2;\ a, b \in R$

Let $c \in R$

Consider

$$c\bar{u} = c(a\bar{v}_1 + b\bar{v}_2)$$
$$= c(a\bar{v}_1) + c(b\bar{v}_2)$$
$$= ca\bar{v}_1 + cb\bar{v}_2$$

Here $c, a, b \in R \Longrightarrow ca, cb \in R$

$\therefore c\bar{u} \in H$ - (3)

From (1), (2) and (3) H is subspace of V

Problem:

Q. **Let w be the set of all vector of the form** $\begin{bmatrix} 5b+2c \\ b \\ c \end{bmatrix}$ **where b and c are arbitrary find vector** $\bar{u}$ **and** $\bar{v}$ **such that w = span** $\{\bar{u}, \bar{v}\}$ **is w a vector space.**

Sol. Given

$$\begin{bmatrix} 5b+2c \\ b \\ c \end{bmatrix} = b\begin{bmatrix} 5 \\ 1 \\ 0 \end{bmatrix} + c\begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$$

$$\bar{v}_1 = \begin{bmatrix} 5 \\ 1 \\ 0 \end{bmatrix}, \qquad \bar{v}_2 = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$$

w= span $\{\overline{v_1}, \overline{v_2}\}$

By the theorem w is subspace

The Null space of a Matrix

Consider the following system of Homogeneous equation of

$$2x + 4y = 0$$
$$3x + 5y = 0$$

The matrix equation is $Ax = B$

A= $\begin{bmatrix} 2 & 4 \\ 3 & 5 \end{bmatrix}$ x= $\begin{bmatrix} x \\ y \end{bmatrix}$ B= $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$\bar{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ is solution of given Homogeneous equation.

The solution set is not empty

We call the set x satisfying $A\bar{x}=\bar{0}$ the null space of A

Definition: Null space of a matrix

The Null space of a m X n matrix A written as Null A is the set of all solutions to the Homogeneous equations $A\bar{x}=\bar{0}$ we write Null A=$\{\bar{x}: \bar{x} \in R^{n\times n}$ and $A\bar{x} = \bar{0}\}$

1Q) Let A= $\begin{bmatrix} 1 & -3 & -2 \\ -5 & 9 & 1 \end{bmatrix}$ **and** $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \bar{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Let $\bar{u} = \begin{bmatrix} 5 \\ 3 \\ -2 \end{bmatrix}$ **determine it $\bar{u}$ belongs to the null space a and find the null space of A.**

Sol: Given

$$A=\begin{bmatrix} 1 & -3 & -2 \\ -5 & 9 & 1 \end{bmatrix} \qquad \bar{u} = \begin{bmatrix} 5 \\ 3 \\ -2 \end{bmatrix}$$

We have $A\bar{x} = \bar{0}$

Consider

$$A\bar{u} = \begin{bmatrix} 1 & -3 & -2 \\ -5 & 9 & 1 \end{bmatrix}\begin{bmatrix} 5 \\ 3 \\ -2 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & -9 & +4 \\ -25 & 27 & -2 \end{bmatrix}$$

$$= \begin{bmatrix} 9-9 \\ 27-27 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Therefore $\bar{u}$ belongs to nullspace A

$\therefore$ $\bar{u} \in$ Nullspace A.

2Q) **A=** $\begin{bmatrix} \mathbf{3} & \mathbf{-5} & \mathbf{-3} \\ \mathbf{6} & \mathbf{-2} & \mathbf{0} \\ \mathbf{-8} & \mathbf{4} & \mathbf{1} \end{bmatrix}$ **find whether** $\bar{w}=$ $\begin{bmatrix} \mathbf{1} \\ \mathbf{3} \\ \mathbf{-4} \end{bmatrix}$ **belongs to null A. Determine the nullspace of A.**

Sol. Given A= $\begin{bmatrix} 3 & -5 & -3 \\ 6 & -2 & 0 \\ -8 & 4 & 1 \end{bmatrix}$ $\bar{w}=\begin{bmatrix} 1 \\ 3 \\ -4 \end{bmatrix}$

We have $A\bar{x} = \bar{0}$

Consider

$$A\bar{w} = \begin{bmatrix} 3 & -5 & -3 \\ 6 & -2 & 0 \\ -8 & 4 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \\ -4 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & -15 & +12 \\ 6 & -6 & 0 \\ -8 & +12 & -4 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Therefore $\bar{w} \in$ null A

<u>Null Space:</u>

A= $\begin{bmatrix} 3 & -5 & -3 \\ 6 & -2 & 0 \\ -8 & 4 & 1 \end{bmatrix}$ $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ $\bar{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

$3x_1 - 5x_2 - 3x_3 = 0$- (1)

$6x_1 - 2x_2 - 0x_3 = 0$- (2)

$-8x_1 + 4x_2 + 1x_3 = 0$-(3)

3x(3) $-24x_1 - 12x_2 - 3x_3 = 0$

$3x_1 - 5x_2 - 3x_3 = 0$

$-21x_1 - 7x_2 = 0$ - (4)

(2) $6x_1 = 2x_2$

$x_2 = 3x_1$

(1) $3x_1 - 5x_2 - 3x_3 = 0$

$3x_1 - 5(3x_1) - 3x_3 = 0$

$3x_1 - 15x_1 - 3x_3 = 0$

$12x_1 = 3x_3$

$x_3 = -4x_1$

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x \\ 3x_1 \\ -4x_1 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 3 \\ -4 \end{bmatrix}$$

$$K \begin{bmatrix} 1 \\ 3 \\ -4 \end{bmatrix}$$

Null space of A= K$\begin{bmatrix} 1 \\ 3 \\ -4 \end{bmatrix}$

Null A= span $\{\bar{w}\}$

Theorem:

The null space of an mxn matrix is a subspace of R^n equivalently the set of all solutions to a system A$\bar{x} = \bar{0}$, of m homogeneous linear equation in n unknowns is a subspace of R^n

Proof: Given A is mXn matrix

We known that null space of A= $\{\bar{w}: A\bar{w} = \bar{0}\}$

To prove that Null space is subspace of R^n

1) Clearly A$\bar{0} = \bar{0}$

$\therefore\ \bar{0}\epsilon$ Null space

2) Let $\bar{u}, \bar{v} \in$ Null space of A

$\Longrightarrow A\bar{u} = \bar{0}andA\bar{v} = \bar{0}$ – (1)

Consider A$(\bar{u} + \bar{v}) = A\bar{u} + A\bar{v}$

= $\bar{0} + \bar{0}$ (from (1))

= $\bar{0}$

A$(\bar{u} + \bar{v})$=$\bar{0}$

$\Longrightarrow \bar{u} + \bar{v} \in$null space of A

3) Consider $A(c\bar{u})$ = $c(A\bar{u})$

= $c(\bar{0})$

= $\bar{0}$

$A(c\bar{u})$ = 0

$c\bar{u}$ ∈null space of A

Therefore Null space of A is subspace of R^n

Hence the theorem

Q. Determine whether the set w=$\left\{\begin{bmatrix} \boldsymbol{b-5d} \\ \boldsymbol{2b} \\ \boldsymbol{2d+1} \\ \boldsymbol{d} \end{bmatrix} \boldsymbol{b,d \in R}\right\}$ is a vector space and if so find the spanning set of N.

Sol: Given w=$\left\{\begin{bmatrix} b-5d \\ 2b \\ 2d+1 \\ d \end{bmatrix} b,d \in R\right\}$

If w is a vectorspace then $\bar{0} \in w$ for some b,d $\begin{bmatrix} b-5d \\ 2b \\ 2d+1 \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

$b-5d = 0$

$2b = 0 \implies b = 0$

$2d+1 = 0 \implies d = -\frac{1}{2}$

$d = 0 \implies d = 0$

$d = 0, -\frac{1}{2}$ *is not possible*

$\therefore \bar{0} \epsilon w$

$\implies w$ *is not a vector space*

Q) $w = \left\{ \begin{bmatrix} r \\ s \\ t \end{bmatrix} : SR - 1 = St2t; S, t \in R \; r, s, t \in R \right\}$

Sol: $w = \left\{ \begin{bmatrix} r \\ s \\ t \end{bmatrix} : SR - 1 = St2t; S, t \in R \right\}$

$For\ some\ r, s, t \begin{bmatrix} r \\ s \\ t \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

$r = 0$

$s = 0$

$t = 0$

$5R - 1 = S + 2t$

$5(0) - 1 = 0 + 2(0)$

$-1 \neq 0$

$\bar{0} \notin w$

$\Longrightarrow w\ is\ not\ a\ vector\ space.$

Q) Determine whether the set w= $\left\{ \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \\ \boldsymbol{d} \end{bmatrix} : \boldsymbol{a - 2b = 4c, 2a = c + 3d} \right\}$ **is a vector space and it so find the spanning set.**

Sol: Given w= $\left\{ \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} : a - 2b = 4c, 2a = c + 3d \right\}$

The set of vector in w are nothing but the solutions of the system of equations

$a - 2b = 4c$

$2a = c + 3d$

$\Longrightarrow a - 2b - 4c = 0$

$2a - c - 3d = 0$

This is system of homogeneous equations and we know that null A is vectorspace and subspace of R^n

$$A = \begin{bmatrix} 1 & -2 & -4 & 0 \\ 2 & 0 & -1 & -3 \end{bmatrix}_{2\times 4\ matrix}$$

$W\ is\ subspace\ of\ R^n$

Solving (1) and (2)

$2\ x\ (1)\ 2a - 4b - 8c\ =\ 0$

$2a - c - 3d = 0$

$-\quad +\quad +$

$-4b - 7c + 3d = 0$

$3d \quad = \quad 4b\ +\ 7c$

$d \quad = \quad \frac{4}{3}b + \frac{7}{3}c$

$a =\ 2b + 4c$

$b = b$

$c = c$

$d = \frac{4}{3}b + \frac{7}{3}c$

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = b\begin{bmatrix} 2 \\ 1 \\ 0 \\ 4/3 \end{bmatrix} + c\begin{bmatrix} 4 \\ 0 \\ 1 \\ 7/3 \end{bmatrix}$$

$$= \frac{1}{3}b\begin{bmatrix} 6 \\ 3 \\ 0 \\ 4 \end{bmatrix} + \frac{1c}{3}\begin{bmatrix} 12 \\ 0 \\ 3 \\ 7 \end{bmatrix}$$

$=k\bar{u} + l\bar{v}$

$\therefore w = span\{\bar{u}, \bar{v}\}$

Q) **If w=** $\left\{\begin{bmatrix} a \\ b \\ c \end{bmatrix} : a - 3b - c = 0\right\}$ **show that w is a subspace of** R^3

Sol: we know that nul A is subspace of R^n.

$Here\ n = 3\ therefore\ w\ is\ subspace\ of\ R^3$

$Given\ a - 3b - c = 0$

$\Longrightarrow\ a = 3b + c;\ b = b;\ c = c$

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = b\begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix} + c\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

$$= \quad k\bar{u} + l\bar{v}$$

$$W \quad = \quad span\ \{\bar{u}, \bar{v}\}$$

The Column space of a matrix

The column space of an mxn matrix A written as col A is the set of all linear combinations of the columns of A

$$If\ A = \{\bar{a}_1, \bar{a}_2, \dots\dots\dots, \bar{a}_n\}\ then$$

$$Col\ A = span\ \{\bar{a}_1, \bar{a}_2, \dots\dots\dots, \bar{a}_n\}$$

Note:

1) We know that span of finite number of vectors is a subspace therefore col A is also a subspace

2) The column space of an mXn matrix A is a subspace of R^m

1Q) **Let A=** $\begin{bmatrix} 7 & -2 & 0 \\ -2 & 0 & 5 \\ 0 & -5 & 7 \\ -5 & 7 & -2 \end{bmatrix}_{4\ X\ 3}$ **if the column space of A is subspace of** R^k

find k

Sol: Here A $is\ a\ 4X3$ matrix

$m = 4, n = 3$

Col a is subspace of R^4

$i.e\ \ K = 4$

2Q) If the null space of A is subspace of R^k what is k

Sol: Here a is a 4x3 matrix

$M = 4, n = 3$

∴Null spae is subspace of R^3

<u>The Matrix Equation $A\bar{x} = \bar{b}$</u>

Let A = $\begin{bmatrix} 1 & 2 & -1 \\ 0 & -5 & 3 \end{bmatrix}$ $\bar{x} = \begin{bmatrix} 4 \\ 3 \\ 7 \end{bmatrix}$

Consider $A\bar{x} = \begin{bmatrix} 1 & 2 & -1 \\ 0 & -5 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \\ 7 \end{bmatrix}$

$$\begin{bmatrix} 4 & +6 & -7 \\ 0 & -15 & +21 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix} = \bar{b}$$

The same result can be achieved in a different way as follows.

$$A\bar{x} = \begin{bmatrix} 1 & 2 & -1 \\ 0 & -5 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \\ 7 \end{bmatrix}$$

$$= 4\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3\begin{bmatrix} 2 \\ -5 \end{bmatrix} + 7\begin{bmatrix} -1 \\ 3 \end{bmatrix}$$

$$= \begin{bmatrix} 4 \\ 0 \end{bmatrix} + \begin{bmatrix} 6 \\ -15 \end{bmatrix} + \begin{bmatrix} -7 \\ 21 \end{bmatrix}$$

$$= \begin{bmatrix} 3 \\ 6 \end{bmatrix} = \bar{b}$$

In the second method the columns of A are multiplied with elements of $\bar{x}$. In other words the multiplication of $A\bar{x}$ is regarded as linear combination of columns of A. Clearly $A\bar{x}$ is defined only if the number of columns of a is equal to the number of rows in $\bar{x}$.

NOTE: Col A=$\{\bar{b}: \bar{b} = A\bar{x}; \bar{x} \in col\ R^n\}$

Problems:

Q) Find a matrix a such that w= Col a where w= $\left\{\begin{bmatrix} \boldsymbol{2r+3t} \\ \boldsymbol{r+s-2t} \\ \boldsymbol{4r+s} \\ \boldsymbol{3r-s-t} \end{bmatrix} : \boldsymbol{r,s,t \in R}\right\}$

Sol: Given w = $\left\{\begin{bmatrix} 2r+3t \\ r+s-2t \\ 4r+s \\ 3r-s-t \end{bmatrix} : r,s,t \in R\right\}$

$$r\begin{bmatrix}2\\1\\4\\3\end{bmatrix} + s\begin{bmatrix}0\\1\\1\\-1\end{bmatrix} + t\begin{bmatrix}3\\-2\\0\\1\end{bmatrix}$$

Col A = $\begin{bmatrix} 2 & 0 & 3 \\ 1 & 1 & -2 \\ 4 & 1 & 0 \\ 3 & -1 & -1 \end{bmatrix}$

W is nothing but sot of all linear combinations of columns of A

Q) Find a matrix A such that w= Col A where w= $\left\{\begin{bmatrix} \boldsymbol{6a-b} \\ \boldsymbol{a+b} \\ \boldsymbol{-7a} \end{bmatrix} : \boldsymbol{a,b \in R}\right\}$

Sol: Given w= $\left\{\begin{bmatrix} 6a-b \\ a+b \\ -7a \end{bmatrix} : a,b \in R\right\}$

$$a\begin{bmatrix}6\\1\\7\end{bmatrix} + b\begin{bmatrix}-1\\1\\0\end{bmatrix}$$

Col A = $\begin{bmatrix} 6 & -1 \\ 1 & 1 \\ -7 & 0 \end{bmatrix}$

W is nothing but set of all linear combination of columns of A.

3Q. Find a matrix A such that w= Col A where w= $\left\{\begin{bmatrix} \boldsymbol{b-c} \\ \boldsymbol{2b+c+d} \\ \boldsymbol{5c-d} \\ \boldsymbol{d} \end{bmatrix}; \boldsymbol{b,c,d \in R}\right\}$

Sol: Given w= $\left\{\begin{bmatrix} b-c \\ 2b+c+d \\ 5c-d \\ d \end{bmatrix}; b,c,d \in R\right\}$

$$b\begin{bmatrix}1\\2\\0\\0\end{bmatrix}+c\begin{bmatrix}-1\\1\\5\\0\end{bmatrix}+d\begin{bmatrix}0\\1\\-1\\1\end{bmatrix}$$

$$\text{Col A}=\begin{bmatrix}1&-1&0\\2&1&1\\0&5&-1\\0&0&1\end{bmatrix}$$

W is nothing but set of all linear combinations of columns of A

3Q. Find a matrix A such that w= Col A where w= $\left\{\begin{bmatrix}\boldsymbol{b-c}\\\boldsymbol{2b+c+d}\\\boldsymbol{5c-d}\\\boldsymbol{d}\end{bmatrix};\boldsymbol{b,c,d}\in\boldsymbol{R}\right\}$

Sol: Given w= $\left\{\begin{bmatrix}b-c\\2b+c+d\\5c-d\\d\end{bmatrix};b,c,d\in R\right\}$

$$b\begin{bmatrix}1\\2\\0\\0\end{bmatrix}+c\begin{bmatrix}-1\\1\\5\\0\end{bmatrix}+d\begin{bmatrix}0\\1\\-1\\1\end{bmatrix}$$

$$\text{Col A}=\begin{bmatrix}1&-1&0\\2&1&1\\0&5&-1\\0&0&1\end{bmatrix}$$

W is nothing but set of all linear combinations of columns of A

Q) Find k when

1) Col A is sunspace of R^k

2) Nul A is subspace of R^k

1) $\begin{bmatrix}2&4&-2&1\\-2&-5&7&3\\3&7&-8&6\end{bmatrix}$

2) $\begin{bmatrix}2&-6\\-1&3\\-4&12\\3&-9\end{bmatrix}$ 3) $\begin{bmatrix}4&5&-2&6&0\\1&1&0&1&0\end{bmatrix}$

1) Here m=3, n=4

$w.k.t$ null space of A is subspace of R^n

∴null space is subspace of R^4

w.k.t Col space is subspace of R^m

∴Col space is subspace of R^3

2) Here m=4, n=2

w.k.t null space of A is subspace of R^n

∴null space is subspace of R^2

w.k.t Col space is subspace of R^m

∴Col space is subspace of R^4

3) Here m=2, n=5

w.k.t null space of A is subspace of R^n

∴null space is subspace of R^5

w.k.t Col space is subspace of R^m

∴Col space is subspace of R^2

Q) **If A=** $\begin{bmatrix} \mathbf{2} & \mathbf{-6} \\ \mathbf{-1} & \mathbf{3} \\ \mathbf{-4} & \mathbf{12} \\ \mathbf{3} & \mathbf{-9} \end{bmatrix}$ **and a non- zero vector in col a & a non zero vector in null A**

Sol: A= $\begin{bmatrix} 2 & -6 \\ -1 & 3 \\ -4 & 12 \\ 3 & -9 \end{bmatrix}$

$w.k.t$ col A is linear combination of column of A each column of A each column of A is an element of Col A.

$\begin{bmatrix} 2 \\ -1 \\ 4 \\ 3 \end{bmatrix} \begin{bmatrix} -6 \\ 3 \\ 12 \\ 9 \end{bmatrix}$ are non zero vectors of col A

To find a non zero vector in nul A $\bar{x}$ such that, $A\bar{x} = \bar{0}$

i.e = $\begin{bmatrix} 2 & -6 & 0 \\ -1 & 3 & 0 \\ 4 & 12 & 0 \\ 3 & -9 & 0 \end{bmatrix}$

$$R_1 \to R_2$$

$$\begin{bmatrix} -1 & 3 & 0 \\ 2 & 3 & 0 \\ 4 & 12 & 0 \\ 3 & -9 & 0 \end{bmatrix}$$

$$R_1 \to R_2 + 2R_1, |R_3 \to R_3 - 4R_1 | R_4 \to R_4 + 3R_1$$

$$\begin{bmatrix} -1 & 3 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Then the equations are $A\bar{x} = \bar{0}$

$$\Longrightarrow -x_1 + 3x_2 = 0$$

$x = 3x_2$

$\Longrightarrow x_2 = x_2$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3x_2 \\ x_2 \end{bmatrix} = x_2 \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

∴A non zero vector in null A = $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$

Q) **A=** $\begin{bmatrix} \mathbf{2} & \mathbf{4} & \mathbf{-2} & \mathbf{1} \\ \mathbf{-2} & \mathbf{-5} & \mathbf{7} & \mathbf{3} \\ \mathbf{3} & \mathbf{7} & \mathbf{-8} & \mathbf{6} \end{bmatrix}$ **find non zero vector in col A & non- zero in null A.**

Sol: A= $\begin{bmatrix} 2 & 4 & -2 & 1 \\ -2 & -5 & 7 & 3 \\ 3 & 7 & -8 & 6 \end{bmatrix}$

w.k.t col A is linear combination of column of A each column of A each column of A is an element of Col A.

$\begin{bmatrix} 2 \\ -2 \\ 3 \end{bmatrix}\begin{bmatrix} 4 \\ -5 \\ 7 \end{bmatrix}\begin{bmatrix} -2 \\ 7 \\ -8 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \\ 6 \end{bmatrix}$ are the non- zero vectors of Col A

Find non zero vector in null A

$$A\bar{x} = \bar{0}$$

$$\begin{bmatrix} 2 & 4 & -2 & 1 & 0 \\ -2 & -5 & 7 & 3 & 0 \\ 3 & 7 & -8 & 6 & 0 \end{bmatrix}$$

$$R_1 \to R_2 + R_1$$

$$\begin{bmatrix} 2 & 4 & -2 & 1 & 0 \\ 0 & 1 & 5 & 4 & 0 \\ 3 & 7 & -8 & 6 & 0 \end{bmatrix}$$

$$R_3 \to 2R_3 + 3R_1$$

$$\begin{bmatrix} 2 & 4 & -2 & 1 & 0 \\ 0 & -1 & 5 & 4 & 0 \\ 3 & 7 & -10 & 9 & 0 \end{bmatrix}$$

$$R_3 \to R_3 + 2R_2$$

$$\begin{bmatrix} 2 & 4 & -2 & 1 & 0 \\ 0 & -1 & 5 & 4 & 0 \\ 0 & 0 & 0 & 17 & 0 \end{bmatrix}$$

$$R_1 \to R_1 + 4R_2, |R_3 \to R_3|17$$

$$\begin{bmatrix} 2 & 0 & 18 & 17 & 0 \\ 0 & -1 & 5 & 4 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$R_1 \to R_1 - 17\ |R_3 \to R_2 - 4R_3|$$

$$\begin{bmatrix} 2 & 0 & 18 & 0 & 0 \\ 0 & -1 & 5 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$R_1 \to \frac{R_1}{2}|R_3 \to R_3| - 1$$

$$\begin{bmatrix} 1 & 0 & 9 & 0 & 0 \\ 0 & 1 & -5 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The equations are $x_1 + 0x_2 + 19x_3 + 0x_4 = 0$

$$0x_1 + x_2 + 5x_3 + 0x_4 = 0$$
$$0x_4 = 0$$

$$\Longrightarrow x_1 + 9x_3 = 0$$
$$x_2 - 5x_3 = 0$$
$$x_4 = 0$$
$$\Longrightarrow x_1 = -9x_3$$

$x_2 = 5x_3$

$x_3 = x_3$

$x_4 = 0$

$$\Longrightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -9x_3 \\ 5x_3 \\ x_3 \\ 0 \end{bmatrix} = x_3 \begin{bmatrix} -9 \\ 5 \\ 1 \\ 0 \end{bmatrix}$$

A non- zero vector in null A is $\begin{bmatrix} -9 \\ 5 \\ 1 \\ 0 \end{bmatrix}$

Q) **A=**$\begin{bmatrix} \mathbf{1} & \mathbf{3} & \mathbf{5} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{4} & \mathbf{-2} \end{bmatrix}$

Sol: w.k.t col A is linear combination of column of A each column of A each column of A is an element of Col A.

$\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 5 \\ 4 \end{bmatrix}$are non zero vector of col A

To find non zero vectors in null$A\bar{x} = \bar{0}$

i.e $\begin{bmatrix} 1 & 3 & 5 & 0 & 0 \\ 0 & 1 & 4 & -2 & 0 \end{bmatrix}$

$$R_1 \longrightarrow R_1 - 3R$$

$$\begin{bmatrix} 1 & 0 & -7 & 6 & 0 \\ 0 & 1 & 4 & -2 & 0 \end{bmatrix}$$

$$\Longrightarrow x_1 - 7x_3 + x_4 = 0$$

$$x_1 + 4x_3 - 2x_4 = 0$$

$$\Longrightarrow x_1 = 7x_3 - 6x_4 \mid x_3 = x_3$$

$$x_1 = -4x_3 + 2x_4 \quad \mid x_4 = x_4$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 7x_3 \\ -4x_3 \\ x_3 \\ 0 \end{bmatrix} + \begin{bmatrix} 6x_4 \\ 2x_4 \\ 0 \\ x_4 \end{bmatrix}$$

$$= x_3 \begin{bmatrix} 7 \\ -4 \\ 1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -6 \\ 2 \\ 0 \\ 1 \end{bmatrix}$$

$\begin{bmatrix} 7 \\ -4 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -6 \\ 2 \\ 0 \\ 1 \end{bmatrix}$ are non- zero vector of null A.

Linear Transformations

Definition:

A transformation or (a function) T from R^n *in* R^n is said to be linear if

1) $T(\bar{u} + \bar{v}) = T(\bar{u}) + T(\bar{v}) where\ \bar{u}, \bar{v} \in R^n$.

2) $T(c\,\bar{u}) = c\,T(\overline{u}) \forall\ \bar{u} \in R^n \& c \in R$.

1Q) Verify whatever the following transformations are linear.

$$T \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x + y \\ y \\ x - z \end{bmatrix}$$

b) $T(x, y) = (x + 1, y, x + y)$

a) $T\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x+y \\ y \\ x-z \end{bmatrix}$

$\Rightarrow T : R^3 \rightarrow R^3.$

$\Rightarrow T \; is \; a \; matrix \; from \; R^3 \; to \; R^3$

$$Let \; \bar{u} = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix}, \bar{v} = \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix}$$

$$\Rightarrow T(\bar{u}) = T\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} x_1+y_1 \\ y_1 \\ x_1-z_1 \end{bmatrix}$$

$$\Rightarrow T(\bar{v}) = T\begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} = \begin{bmatrix} x_2+y_2 \\ y_2 \\ x_2-z_2 \end{bmatrix}$$

$$\bar{u}+\bar{v} = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} + \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} = \begin{bmatrix} x+x_2 \\ y_1+y_2 \\ z_1-z_2 \end{bmatrix}$$

$$T(\bar{u}+\bar{v}) = T\begin{bmatrix} x_1+x_2 \\ y_1+y_2 \\ z_1-z_2 \end{bmatrix} = \begin{bmatrix} x_1+x_2+y_1+y_2 \\ y_1+y_2 \\ x_1+x_2-(z_1+z_2) \end{bmatrix}$$

$$= \begin{bmatrix} (x_1+y_1)+(x_2+y_2) \\ y_1+y_2 \\ (x_1-z_1)+(x_2-z_2) \end{bmatrix}$$

$$= \begin{bmatrix} x_1+y_1 \\ y_1 \\ x_1-z_1 \end{bmatrix} + \begin{bmatrix} x_2+y_2 \\ y_2 \\ x_2-z_2 \end{bmatrix}$$

$$= T(\bar{u}) + T(\bar{v})$$

$Lert \; C \; be \; any \; scalar$

To verify $T(c\bar{u}) = cT\,(\bar{u})$

$$C\,\bar{u} = c\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} c\,x_1 \\ c\,y_1 \\ c\,z_1 \end{bmatrix}$$

$$= T(c\bar{u}) = T\begin{bmatrix} c\,x_1 \\ c\,y_1 \\ c\,z_1 \end{bmatrix} = \begin{bmatrix} cx_1 + cy_1 \\ c\,y_1 \\ cx_1 - c\,z_1 \end{bmatrix}$$

$$= c\begin{bmatrix} x_1 + y_1 \\ y_1 \\ x_1 - z_1 \end{bmatrix} = C\,T\,(\bar{u})$$

$\therefore$ *is a linear Transformation*

2Q) Given $T(x, y) = (x + 1, y, x + y)$

Here $T: R^2 \to R^3$

$let\ \bar{u} = (x_1, y_1), \bar{v} = (x_2, y_2)$

$T(\bar{u}) = (x_1 + 1, y_1, x_1 + y_1)$

$T(\bar{v}) = (x_2 + 1, y_2, x_2 + y_2)$

$\bar{u} + \bar{v} = (x_1 + x_2, y_1 + y_2)$

$T(\bar{u} + \bar{v}) = T\,(x_1 + x_2, y_1 + y_2) = (x_1 + x_2 + 1, y_1 + y_2, (x_1 + y_2) + (y_2 + y_2)$

$Consider\ T(\bar{u}) + T(\bar{v})$

$= (x_1 + 1, y_1, x_1 + y_1) + (x_2 + 1, y_2, x_2 + y_2)$

$= (x_1 + x_2 + 2, y_1 + y_2, (x_1 + y_1) + (x_2 + y_2))$

$T(\bar{u} + \bar{v}) = \neq T(\bar{u}) + T(\bar{v})$.

$T\ is\ not\ linear\ transformation.$

3Q) **Let A=** $\begin{bmatrix} \mathbf{2} & \mathbf{0} \\ \mathbf{0} & \mathbf{2} \end{bmatrix}$ **and define T: R²→R² by** $\boldsymbol{T(\bar{x}) = A\bar{x}}$ **. Find the images of**

$$\bar{\boldsymbol{u}} = \begin{bmatrix} \mathbf{1} \\ \mathbf{-3} \end{bmatrix} \bar{\boldsymbol{v}} = \begin{bmatrix} \mathbf{0} \\ \mathbf{-4} \end{bmatrix} \bar{\boldsymbol{w}} = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}$$

Sol: Given A =$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ Also given $T(\bar{x}) = A\bar{x}$

$$T(\bar{u}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \\ -6 \end{bmatrix}$$

$$T(\bar{v}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 0 \\ -4 \end{bmatrix} = \begin{bmatrix} 0 \\ -8 \end{bmatrix}$$

$$T(\bar{w}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2a \\ 2b \end{bmatrix}$$

In general if $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ then $T(\bar{x}) = A\bar{x}$

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = 2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

4Q) **Let A=** $\begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{-2} \\ \mathbf{-2} & \mathbf{1} & \mathbf{6} \\ \mathbf{3} & \mathbf{-2} & \mathbf{-5} \end{bmatrix}, \bar{\boldsymbol{b}} = \begin{bmatrix} \mathbf{-1} \\ \mathbf{7} \\ \mathbf{-3} \end{bmatrix}$

Define T: $R^3 \rightarrow R^3$ by $T(\bar{x}) = A\bar{x}$. Find α vector $\bar{x}$ whose image under T is $\bar{b}$ also determine whether $\bar{x}$ is unique

Sol: We have to find $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \exists\ T(\bar{x}) = \bar{b}$ i.e $A\bar{x}=\bar{b}$

We have to solve the matrix equation

$$A\bar{x} = \bar{b} \text{ i.e } \begin{bmatrix} 1 & 0 & -2 \\ -2 & 1 & 6 \\ 3 & -2 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 7 \\ -3 \end{bmatrix}$$

Consider the augmented matrix

$$\left[A/\bar{b} \right] = \begin{bmatrix} 1 & 0 & -2 & -1 \\ -2 & 1 & 6 & 7 \\ 3 & -2 & -5 & -3 \end{bmatrix}$$

$$R_2 \rightarrow R_2 + 2R_1, R_3 \longrightarrow R_3 - 3R_1$$

$$\begin{bmatrix} 1 & 0 & -2 & -1 \\ 0 & 1 & 2 & 5 \\ 0 & -2 & 1 & 0 \end{bmatrix}$$

$$R_3 \longrightarrow R_3 + 2R_2$$

$$\begin{bmatrix} 1 & 0 & -2 & -1 \\ 0 & 1 & 2 & 5 \\ 0 & 0 & 5 & 10 \end{bmatrix}$$

$$R_3 \longrightarrow R_3/5$$

$$\begin{bmatrix} 1 & 0 & -2 & -1 \\ 0 & 1 & 2 & 5 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

$$R_1 \longrightarrow R_1 + 2R_3\ , R_2 \rightarrow R_2 - 2R_2$$

$$\begin{bmatrix} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

The number of pivot columns =3 and no. of variables=3 Therefore the system is consistent and has unique solution and the solution is $\bar{x} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$

Verification

$\begin{bmatrix} 1 & 0 & -2 \\ -2 & 1 & 6 \\ 3 & -2 & -5 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3-4 \\ -6+1+12 \\ 9-2-10 \end{bmatrix} = \begin{bmatrix} -1 \\ 7 \\ -3 \end{bmatrix}$ Hence verified

Q) **A=**$\begin{bmatrix} \mathbf{1} & \mathbf{-5} & \mathbf{-7} \\ \mathbf{-3} & \mathbf{7} & \mathbf{5} \end{bmatrix} \bar{\boldsymbol{b}} = \begin{bmatrix} \mathbf{-2} \\ \mathbf{-2} \end{bmatrix}$

Sol: We have to find $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = T\bar{x} = \bar{b}$ i.e $A\bar{x} = \bar{b}$We have to solve the matrix equation $A\bar{x} = \bar{b}$ i.e

$$\begin{bmatrix} 1 & -5 & -7 \\ -3 & 7 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$$

Consider the augmented matrix

$$\left[A/\bar{b} \right] = \begin{bmatrix} 1 & -5 & -7 & -2 \\ -3 & 7 & 5 & -2 \end{bmatrix}$$

$$R_2 \to R_2 + 3R_1$$

$$\begin{bmatrix} 1 & -5 & -7 & -2 \\ 0 & -8 & -16 & -8 \end{bmatrix}$$

$$R_2 \to R_2/-8$$

$$\begin{bmatrix} 1 & -5 & -7 & -2 \\ 0 & 1 & 2 & 1 \end{bmatrix} -2+5$$

$$R_1 \longrightarrow R_1 + 5R_2$$

$$\begin{bmatrix} 1 & 0 & 3 & 3 \\ 0 & 1 & 2 & 1 \end{bmatrix}$$

No.of pivot columns=2 , No of variables =3 The system is consistent and has infinite solution The equations are

$$x_1 + 0x_2 + 3x_3 = 3$$

$$0x_1 + x_2 + 2x_3 = 1$$

$$x_1 + 3x_3 = 3$$

$$x_2 + 2x_3 = 1$$

$$x_1 = -3x_3 + 3$$

$$x_2 = -2x_3 + 1$$

$$x_3 = x_3$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_3 \begin{bmatrix} -3 \\ -2 \\ 1 \end{bmatrix} + \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix}$$

We get infinitely many solutions by receiving various values to x_3

Definition:

$$\boldsymbol{e_1} = (\mathbf{1}, \mathbf{0})$$

$$\boldsymbol{e_2} = (\mathbf{0}, \mathbf{1})$$

The matrix $[T\bar{e_1}, T\bar{e_2} \ldots\ldots]$ is called the standard matrix of T defined on R^n

Problems: Let $T: R^2 \to R^4$ be linear transformation with $T\bar{e_1} = (3,1,3,1), T\bar{e_2} = (-5,2,0,0)\bar{e_1} = (1,0)\bar{e_2} = (0,1)$ write down the standard matrix of T and also find the values of T (3,4)& T(-2,1)

Sol: Given $T: R^2 \to R^4, \bar{e_1} = (1,0)\bar{e_2} = (0,1)$

Also given $T\bar{e_1} = (3,1,3,1)$

$$T\bar{e_2} = (-5,2,0,0)$$

The standard matrix of T=A= $[T(\bar{e_1})T(\bar{e_2})]$

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

To find T(3,4)= A(3,4)

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

$$= \begin{bmatrix} 9 - 20 \\ 3 + 8 \\ 9 \\ 3 \end{bmatrix} = \begin{bmatrix} -11 \\ 11 \\ 9 \\ 3 \end{bmatrix}$$

$$To\ find\ T(-2,1) = A(-2,1)$$

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

$$\begin{bmatrix} -6-5 \\ -2+2 \\ -6 \\ -2 \end{bmatrix} = \begin{bmatrix} -11 \\ 0 \\ -6 \\ -2 \end{bmatrix}$$

Q) If T is a linear transformation then

1) $\mathbf{T(\bar{0}) = \bar{0}}$

2) $\boldsymbol{T(c\bar{u} + d\bar{v}) = cT(\bar{u}) + dT(\bar{v})}$

3) $\boldsymbol{T(\bar{u} - \bar{v}) = T(\bar{u}) - T(\bar{v})}$**for all vectors** $\boldsymbol{\bar{u}, \bar{v}}$ **in the domain of T and all scalar s,c d**

Proof: Given T is a linear transformation therefore

$$T(\bar{u} + \bar{v}) = T(\bar{u}) + T(\bar{v}) \dashrightarrow (1)$$

$$T(c\bar{u}) = cT(\bar{u}) \dashrightarrow (2)$$

Where c is scalar $\bar{u}, \bar{v}$ are vectors

To prove that $T(\bar{0}) = \bar{0}$

We have $T(c\bar{u}) = c.T(\bar{u})$

Put c=0

$\Rightarrow T(0.\bar{u})$=0.T$(\bar{u})$

$\Rightarrow T(\bar{0}) = \bar{0}$

To prove that

$T(c\bar{u} + d\bar{v}) = cT(\bar{u}) + dT(\bar{v})$

Consider

$T(c\bar{u} + d\bar{v}) = T(c\bar{u}) + T(d\bar{v})$ From (1)

$= c.T(\bar{u}) + d.T(\bar{v})$From (2)

$$\therefore T(c\bar{u} + d\bar{v}) = c.T(\bar{u}) + d.T(\bar{v})$$

To prove that $T(\bar{u} - \bar{v}) = T(\bar{u}) - T(\bar{v})$

We have $T(c\bar{u} + d\bar{v}) = c.T(\bar{u}) + d.T(\bar{v})$

Put c=1 , d=-1

$$\mathrm{T}(1.\bar{u} + (-1)\bar{u}) = 1.T(\bar{u}) + (-1), (\bar{v})$$

$$\Rightarrow\ T(\bar{u} - \bar{v}) = T(\bar{u}) - T(\bar{v})$$

Theorem: A Transformation T is linear if and only if

$T(c\bar{u} + d\bar{v}) = cT(\bar{u}) + dT(\bar{v})$for all $\bar{u}, \bar{v}$ in the domain of D and for all scalars c,d

Proof :- Given $\bar{u}, \bar{v}$ are vectors in the domain of D and c,d are scalars

U-T-V

Suppose T is a linear transformation to prove that a

$$T(c\bar{u} + d\bar{v}) = cT(\bar{u}) + dT(\bar{v})$$

Given T is a linear transformation

$$\therefore\ T(\bar{u} + \bar{v}) = T(\bar{u}) + T(\bar{v}) \rightarrow (1)$$

$$T(c\bar{u}) = c.T(\bar{u}) \qquad \rightarrow (2)$$

Consider

$$T(c\bar{u} + d\bar{v}) = c.T(\bar{u}) + d.T(\bar{v}) \rightarrow (3)$$

To prove that T is linear

C=1, d=1

$$T(\bar{u} + \bar{v}) = T(\bar{u}) + T(\bar{v})$$

Put c=c, d=0

$$T(c\bar{u} + 0.\bar{v}) = c.T(\bar{u}) + 0.T(\bar{v})$$

$$T(c\bar{u}) = c.T(\bar{u}) + 0$$

$$T(c\bar{u}) = c.T(\bar{u})$$

Therefore T is linear transformation a

<u>Definition :</u>

Let T: V →w be a linear transformation then the kernel(T) is the set of all $\bar{u}\ in\ V$

Such that a $T\ (\bar{u}) = \bar{0}$(The zero vector in w)

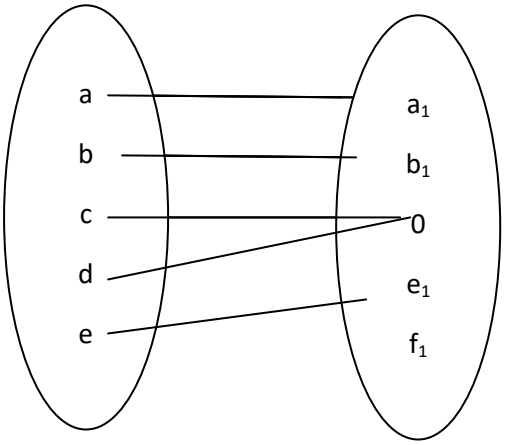
The range of T is the set of all vectors $\bar{w}\exists\ T(\bar{u}) = \bar{w}$

Kernel of T ={c,d}

Domain = v={ a,b,c,d.e}

Co-domain= w={a,b,o,e,f}

Range={a,b,o,e}



T(a)=a_1,T(b)=b_1, T(c)=0, T(d)=0, T(e)=e_1 kernel of T ={c,d}

Domain = v= { a,b,c,d.e}

Co-domain= w={a_1,b_1,o,e_1,f_1}

Range={a_1,b_1,o,e_1}

Theorem : Let T:V→W be a linear transformation then kernel (T) is a subspace of V

Proof: Given T:V→W is a linear transformation

Let $\bar{0} \in V$

We know that $T(\bar{0}) = \bar{0}$

$\Rightarrow \bar{0}$ ∈ kernel of T →(1)

[Kernel of T is set of all elements mapped to $\bar{0}$ i.e ker T=$\{\bar{u}: T(\bar{u}) = \bar{0}\}$

Let $\bar{u}, \bar{v} \in \ker T$

$$\Rightarrow T(\bar{u}) = \bar{0}$$

$$T(\bar{v}) = \bar{0}$$

Consider

$$T(\bar{u} + \bar{v}) = \bar{0}$$

$$\bar{u} + \bar{v} \in kernel\ of\ T \ \rightarrow (2)$$

Let c be any scalar and is kernel of T

$$\Rightarrow T(\bar{u}) = \bar{0}$$

$$\Rightarrow T(c\bar{u}) = c.T(\bar{u})$$

$$= c.\bar{0}$$

$$= \bar{0}$$

$$T(c\bar{u}) = \bar{0}$$

$$c\bar{u} \in kernel\ of\ T \ \rightarrow (3)$$

From (1),(2)&(3) kernel of T is subspace of V

Theorem: Let T:V→W be a linear transformation the range of T is a subspace of W

Proof: Given T:V→W is a linear transformation

To prove that range of T is subspace of W

We know that

$$T(v) = \{T(\bar{x}) : \bar{x} \in v\}$$

Clearly $T(V) \sqsubseteq W$

We have T is a linear transformation

$$\therefore T(\bar{0}) = \bar{0}$$

$$T(\bar{0}) \in T(v) \rightarrow (1)$$

$$\bar{0} \in T(v) \rightarrow (1)$$

Let $\bar{u}, \bar{v} \in V$

$$\Rightarrow T(\bar{u}), T(\bar{v}) \in W$$

We have T$(\bar{u} + \bar{v}) \in W$

$$\Rightarrow T(\bar{u}) + T(\bar{v}) \in W$$

$$\bar{u} + \bar{v} \in W \rightarrow \quad (2)$$

Let c be any scalar and $T(\bar{u}) \in W$

Since T is linear transformation

$$\Rightarrow T(c.\bar{u}) = c.T(\bar{u})$$

$$\therefore c.T(\bar{u}) \in W \rightarrow (3)$$

From (1),(2) &(3) The range of T is a subspace of w

5a) We have to find $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \exists T(\bar{x}) = \bar{b}\ A\bar{x} = \bar{b}$

We have to solve matrix $A\bar{x} = \bar{b}$

$$\begin{bmatrix}1 & -3 & 2\\ 0 & 1 & -4\\ 3 & -5 & -9\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}6\\ -7\\ -9\end{bmatrix}$$

The augmented matrix is

$$\begin{bmatrix}1 & -3 & 2 & 6\\ 0 & 1 & -4 & -7\\ 3 & -5 & -9 & -9\end{bmatrix}$$

$$R_3 \to R_3 - 3R_1$$

$$\begin{bmatrix}1 & -3 & 2 & 6\\ 0 & 1 & -4 & -7\\ 0 & 4 & -15 & -27\end{bmatrix}$$

$$R_3 \to R_3 - 4R_2, R_1 \to R_1 + 3R_2$$

$$\begin{bmatrix}1 & 0 & -10 & -15\\ 0 & 1 & -4 & -7\\ 0 & 0 & -1 & 1\end{bmatrix}$$

The no.of pivot columns =3 and no.of various =3

The system is consistent and have unique solution and solution is $\begin{bmatrix}-5\\ -3\\ 1\end{bmatrix}$

Verification $A\bar{x}=\begin{bmatrix}1 & -3 & 2\\ 0 & 1 & -4\\ 3 & -5 & -9\end{bmatrix}\begin{bmatrix}-5\\ -3\\ 1\end{bmatrix}=\begin{bmatrix}6\\ -7\\ 9\end{bmatrix}$

($\therefore$ Hence Verified)

Linearly Independent sets

An indexed set of vectors $\{\bar{v}_1\bar{v}_2\ldots\ldots\ldots\bar{v}_n\}$ is a vector space v is said to be linearly independent if the vector equation $c_1\bar{v}_1 + c_2\bar{v}_2 + \cdots \ldots\ldots + c_n\bar{v}_n = \bar{0}$ has only the trivial solution $c_1 = 0, c_2 = 0, \ldots \ldots . c_n = 0$

Linear Dependent Sets.

The Set $\overline{v_1}, \overline{v_2} \ldots \ldots \ldots \overline{v_n}$ is said to be linearly dependent if $c_1\bar{v}_1 + c_2\bar{v}_2 +$ $\cdots \ldots \ldots \ldots + c_n\bar{v}_n = 0$ has a non – trivial solution i.e. there are some scalars $c_1, c_2 \ldots \ldots c_n$ not all zero such that $c_1\bar{v}_1 + c_2\bar{v}_2 + \cdots \ldots \ldots + c_n\bar{v}_n = \bar{0}$

Note:

1) A set containing only one non- zero vector is always linearly independent

$\overline{v_1} \neq 0$

Let $c_1\bar{v}_1 = \bar{0}$

$\Longrightarrow c_1 = 0 \text{or} \bar{v}_1 = 0$

Here $\overline{v_1} \neq 0$

$$\therefore c_1 = 0$$

$\therefore \{\bar{v}_1\}$is linearly independent

2) The Set Containing only the zero vectors is always linearly dependent

Let w= $\{\bar{0}\}$

Consider $c_1\bar{0} = \bar{0}$

Here $c_1 \neq 0$

∴w= $\{\bar{0}\}$ is linearly dependent

3) Any set of vectors which include zero vector is always linearly dependent.

H= $\{\bar{0}, \overline{v_1}, \overline{v_2}, \overline{v_3}, \}$

Choose scalars 1, 0, 0, 0

Then $1.\bar{0}_1 + 0\bar{v}_1 + 0.\bar{v}_2 + 0\bar{v}_3 = \bar{0}$ *

Here H is linearly dependent.

4) A set of nonzero vectors is L.D if and only if one vector is a scalar multiple of the other

Let H= $\{\bar{v}_1, \bar{v}_2\}$

Suppose that H is L.D

$\Longrightarrow$ There exists scalars c_1, c_2 not all zero such that $c_1\bar{v}_1 + c_2\bar{v}_2 = \bar{0}$

Suppose $c_1 \neq 0$

$$c_1\bar{v}_1 + c_2\bar{v}_2 = \bar{0}$$

$$c_1\bar{v}_1 = -c_2\bar{v}_2$$

$$\Longrightarrow \bar{v}_1 = \frac{-c_2\bar{v}_2}{c_1}$$

Here $\bar{v}_1$ is scalar multiple of $\bar{v}_2$

Conversely, suppose that one vector is multiple of the other

Let $\bar{v}_1 = c_1\bar{v}_2$

$$\Longrightarrow \bar{v}_1 = c_1\bar{v}_1 = \bar{0}$$

Therefore $\{\bar{v}_1, \bar{v}_2\}$ is linearly dependent

Problems:

1Q. Determinate whether the vectors are linearly independent or linearly depend

$\begin{bmatrix}-1\\1\\0\end{bmatrix}\begin{bmatrix}-2\\0\\1\end{bmatrix}$

Sol: Let $\overline{v_1} = \begin{bmatrix}-1\\1\\0\end{bmatrix}, \overline{v_2} = \begin{bmatrix}-2\\0\\1\end{bmatrix}$

Let c_1, c_2 be scalars consider

$$c_1\bar{v}_1 + c_2\bar{v}_2 = \bar{0}$$

$$c_1\begin{bmatrix}-1\\1\\0\end{bmatrix} + c_2\begin{bmatrix}-2\\0\\1\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

$$\Longrightarrow \begin{bmatrix}-c_1 - 2c_2\\c_1 + 0\\0 + c_2\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

$$\Longrightarrow -c_1 - 2c_2 = 0$$

$$c_1 = 0$$

$c_2 = 0$

$c_1 = 0, c_2 = 0$

$\therefore \{\bar{v}_1, \bar{v}_2\}$is linearly independent

2) $\left\{\begin{bmatrix}1\\0\end{bmatrix},\begin{bmatrix}0\\1\end{bmatrix}\right\}$

Sol: Let $\bar{v}_1 = \begin{bmatrix}1\\0\end{bmatrix}\bar{v}_2 = \begin{bmatrix}0\\1\end{bmatrix}$

Let c_1, c_2 be scalars

$$c_1\bar{v}_1 + c_2\bar{v}_2 = \bar{0}$$

$$c_1\begin{bmatrix}1\\0\end{bmatrix} + c_2\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$$

$$\begin{bmatrix}c_1\\c_2\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$$

$c_1 = 0, c_2 = 0$

$\therefore \{\bar{v}_1,\ \bar{v}_2\}$is linearly independent

3) $\left\{\begin{bmatrix}1\\0\\0\end{bmatrix},\begin{bmatrix}0\\1\\0\end{bmatrix},\begin{bmatrix}0\\0\\1\end{bmatrix}\right\}$

Sol: Let $\bar{v}_1 = \begin{bmatrix}1\\0\\0\end{bmatrix}, \bar{v}_2 = \begin{bmatrix}0\\1\\0\end{bmatrix}, \bar{v}_3 = \begin{bmatrix}0\\0\\1\end{bmatrix}$

Let c_1, c_2, c_3 be scalars.

$$c_1\bar{v}_1 + c_2\bar{v}_2 + c_3\bar{v}_3 = 0$$

$$c_1\begin{bmatrix}1\\0\\0\end{bmatrix} + c_2\begin{bmatrix}0\\1\\0\end{bmatrix} + c_3\begin{bmatrix}0\\0\\1\end{bmatrix} = 0$$

$$\begin{bmatrix}c_1\\c_2\\c_3\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

$c_1 = 0, c_2 = 0, c_3 = 0$

$\{v_1, v_2, v_3\} = 0$linearly independent

Q. Determine whether the set of vectors $\overline{v_1}$= $\begin{bmatrix}-1\\0\\-2\end{bmatrix}$, $\overline{v_2}$ = $\begin{bmatrix}3\\2\\-4\end{bmatrix}$ $\overline{v_3}$ = $\begin{bmatrix}-3\\-5\\1\end{bmatrix}$is linearly independent in R^3

Sol: Let c_1, c_2, c_3 be scalars.

$$c_1\bar{v}_1 + c_2\bar{v}_2 + c_3\bar{v}_3 = 0$$

$$\Rightarrow c_1\begin{bmatrix}-1\\0\\-2\end{bmatrix} + c_2\begin{bmatrix}3\\2\\-4\end{bmatrix} + c_3\begin{bmatrix}-3\\-5\\1\end{bmatrix} = 0$$

Consider the augmented matrix

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

$$R_3 \longrightarrow R_3 - 2R_1$$

$$\begin{bmatrix}-1 & 3 & -3 & 0\\0 & 2 & -5 & 0\\0 & -10 & 7 & 0\end{bmatrix}$$

$$R_3 \longrightarrow R_3 + 5R_2$$

$$\begin{bmatrix}-1 & 3 & -3 & 0\\0 & 2 & -5 & 0\\0 & 0 & -18 & 0\end{bmatrix}$$

$$R_3 \longrightarrow R_3/-18$$

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

$$R_1 \longrightarrow R_1 + 3R_3$$

$$R_2 \longrightarrow R_2 + 5R_3$$

$$\begin{bmatrix}-1 & 3 & 0 & 0\\0 & 2 & 0 & 0\\0 & 0 & 1 & 0\end{bmatrix}$$

$$R_2 \longrightarrow R_2/2$$

$$\begin{bmatrix}-1 & 3 & 0 & 0\\0 & 1 & 0 & 0\\0 & 0 & 1 & 0\end{bmatrix}$$

$$R_1 \longrightarrow R_1 - 3R_2$$

$$\begin{bmatrix} -1 & 3 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$R_1 \longrightarrow \frac{R_1}{-1}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

There are three pivot columns and three variables therefore the solution is unique. This is homogeneous system of equations.

Therefore the vectors are linearly independent.

Q. $\overline{v_1} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \overline{v_2} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \overline{v_3} = \begin{bmatrix} -3 \\ 2 \\ -1 \end{bmatrix}$

Sol: Let c_1, c_2, c_3 be scalars.

$$c_1\bar{v}_1 + c_2\bar{v}_2 + c_3\bar{v}_3 = 0$$

$$c_1 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} + c_3 \begin{bmatrix} -3 \\ 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & -3 & 0 \\ 2 & -2 & 2 & 0 \\ -1 & 1 & -1 & 0 \end{bmatrix}$$

$R_2 \longrightarrow R_2 - 2R_1, R_3 \longrightarrow R_3 + R_1$

$$\begin{bmatrix} 1 & 1 & -3 & 0 \\ 0 & -4 & 8 & 0 \\ 0 & 2 & -4 & 0 \end{bmatrix}$$

$$R_2 \longrightarrow {}^{R_2}/_{-4}, R_3 \longrightarrow {}^{R_3}/_{2}$$

$$\begin{bmatrix} 1 & 1 & -3 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 1 & -2 & 0 \end{bmatrix}$$

$$R_3 \longrightarrow R_3 - R_2$$

$$\begin{bmatrix} 1 & 1 & -3 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$R_1 \longrightarrow R_1 - R_2$$

$$\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Here no. of pivot columns = 2 no. of variables= 3.

The system has infinitely many solutions also last row is zero $\Longrightarrow c_3$ is a free variables.

$\therefore$ The vectors are linearly dependent.

Theorem : An indexed set $\{\bar{v}_1, \bar{v}_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots, \bar{v}_n\}$ non zero vectors is linearly dependent $\Leftrightarrow$ same $\bar{v}_j$ with j>1 is linear combination of preceding vectors $\bar{v}_1, \bar{v}_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots, \bar{v}_n$

Proof : Let $\{\bar{v}_1, \bar{v}_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots, \bar{v}_n\}$ be set of vectors to

Prove that (1) is L.D $\Leftrightarrow$ some $\bar{v}_j$ is linear combination of preceding vectors

i.e.,$\bar{v}_j = C_1\bar{v}_1 + C_2\bar{v}_2 + \ldots\ldots\ldots\ldots . C_{j-1}\bar{v}_{j-1} (j = 1)$

and$C_1, C_2 \ldots\ldots\ldots\ldots\ldots\ldots . C_{j-1}$ not all zero

$\Rightarrow C_1\bar{v}_1 + C_2\bar{v}_2 + \ldots\ldots\ldots\ldots . C_{j-1}\bar{v}_{j-1} + (-1)\bar{v}_j = \bar{o}$

$\Rightarrow C_1\bar{v}_1 + C_2\bar{v}_2 + \ldots\ldots\ldots\ldots . C_{j-1}\bar{v}_{j-1} + (-1)\bar{v}_j = ov_j + ov_{j+2} + \ldots\ldots\ldots\ldots\ldots\ldots + o\bar{v}_n$

$= 0$

$\{\bar{v}_1, \bar{v}_2, \bar{v}_3 \ldots\ldots\ldots\ldots\ldots\ldots \bar{v}_{j-1}, \bar{v}_j \ldots\ldots\ldots . \bar{v}_n\}$is L.D

Conversely suppose that (1) is L.D

There exists scalars $C_1, C_2, \ldots\ldots\ldots\ldots C_n$ not all zero

i.e.,$C_1\bar{v}_1 + C_2\bar{v}_2 + \ldots\ldots\ldots\ldots . \quad C_j\bar{v}_j + \ldots\ldots\ldots . C_n\bar{v}_n = \bar{o}$

Let j be the largest subscript for which $C_j \neq 0$

$C_{j+1} = 0, C_{j+2} = 0,$ $\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots C_n = 0$

$(2) \Rightarrow C_1\bar{v}_1 + C_2\bar{v}_2 + \ldots\ldots\ldots\ldots . C_{j-1}\bar{v}_{j-1} + C_j\bar{v}_j = \bar{o}$

$$\Rightarrow C_j\bar{v}_j = -(C_1\bar{v}_1 + C_2\bar{v}_2 + \ldots\ldots\ldots\ldots . C_{j-1}\bar{v}_{j-1})\bar{o}$$

$$\Rightarrow \bar{v}_j = -\left[\frac{C_1}{C_j}\bar{v}_1 + \frac{C_2}{C_j}\bar{v}_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots . + \frac{C_{j-1}}{C_j}\bar{v}_{j-1}\right]$$

$\frac{C_1}{C_j}\bar{v}_1 + \frac{C_2}{C_j}\bar{v}_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots . + \frac{C_{j-1}}{C_j}\bar{v}_{j-1}$

$\Rightarrow \bar{v}_j$is L.c of preceding vectors suppose j=1 then $C_i\bar{v}_i = 5$

Here $C_1 \neq 0$

$\therefore \bar{v}_i = \bar{o}$

Given (1) contains non-zero vectors

This is a contradiction

∴our assumption j=1 is wrong

$\therefore j = 1$

Hence the theorem

Basis:

Let H be a subspace of a vector space v then an indexed subset of vector of H say p= $\{\overline{b_1}, \overline{b_2} \ldots\ldots \overline{b_n}\}$ is called basis for H if 1) β is L.I set .

2) H = Span β i.e., the subspace spanned by βcornered with H

Problem:

Q. Let $\overline{v_1} = \begin{bmatrix} 3 \\ 0 \\ -6 \end{bmatrix}, \overline{v_2} = \begin{bmatrix} -4 \\ 1 \\ 7 \end{bmatrix}, \overline{v_3} = \begin{bmatrix} -2 \\ 1 \\ 5 \end{bmatrix}$Determine if $\{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$ is a basis for R^3

Sol: Let $\beta = \{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$

To show that β is basis we have to check if 1) β is L.I 2) span of β R^3

Let c_1, c_2, c_3 be scalars.

$$c_1\bar{v}_1 + c_2\bar{v}_2 + c_3\bar{v}_3 = \overline{0}$$

$$c_1\begin{bmatrix}3\\0\\-6\end{bmatrix}+c_2\begin{bmatrix}-4\\1\\7\end{bmatrix}+c_3\begin{bmatrix}-2\\1\\5\end{bmatrix}=\bar{0}$$

Consider the augmented matrix.

$$\begin{bmatrix}3 & -4 & -2 & 0\\0 & 1 & 1 & 0\\-6 & 7 & 5 & 0\end{bmatrix}$$

$$R_3 \longrightarrow R_3 + 2R_1$$

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

$$R_3 \longrightarrow R_3 + R_2$$

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

$$R_1 \longrightarrow R_1 + 4R_2$$

$$\begin{bmatrix}3 & 0 & 2 & 0\\0 & 1 & 1 & 0\\0 & 0 & 2 & 0\end{bmatrix}$$

$R_1 \longrightarrow R_1 - 3, R_2 \longrightarrow 2R_2 - R_3$

$$\begin{bmatrix}3 & 0 & 0 & 0\\0 & 2 & 0 & 0\\0 & 0 & 2 & 0\end{bmatrix}$$

No. of pivot columns =1 and no. of variables= 1. Therefore the set is linearly independent.

To verify that span$\{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$ = R^3

Let $\bar{u} = \begin{bmatrix}a\\b\\c\end{bmatrix}$ consider the equation $x\bar{v}_1 + y\bar{v}_2 + z\bar{v}_3 = \bar{u}$

$$\begin{bmatrix}3 & -4 & -2 & a\\0 & 1 & 1 & b\\-6 & 7 & 5 & c\end{bmatrix}$$

$$R_3 \longrightarrow R_3 + 2R_1$$

$$\begin{bmatrix} 3 & -4 & -2 & a \\ 0 & 1 & 1 & b \\ 0 & -1 & 1 & c+2a \end{bmatrix}$$

$$R_3 \longrightarrow R_3 + R_2$$

$$\begin{bmatrix} 3 & -4 & -2 & a \\ 0 & 1 & 1 & b \\ 0 & 0 & 2 & c+2a+b \end{bmatrix}$$

$R_1 \longrightarrow R_1 - R_3, R_2 \longrightarrow 2R_2 - R_3,$

$$\begin{bmatrix} 3 & 0 & 0 & a+4b-c-2a-b \\ 0 & 2 & 0 & 2b-c-2a-b \\ 0 & 0 & 2 & c+2a+b \end{bmatrix}$$

$$\begin{bmatrix} 3 & 0 & 0 & -a+3b-c \\ 0 & 2 & 0 & -2a+b-c \\ 0 & 0 & 2 & 2a+b+ \end{bmatrix}$$

The number of pivot columns=3 and the no. of variable= 3.

Therefore the solution is unique and the solution is 3x = -a+3b-c

X=$\frac{1}{3}[-a+3b-c]$

$$\Longrightarrow 2y = -2a+b-c$$

$\Longrightarrow$ y=$\frac{1}{2}[-2a+b-c]$

2z = 2a+b+c

z=$\frac{1}{2}$ [2a+b+c]

x=$\frac{1}{3}$ [-a+3b-c]

y=$\frac{1}{2}$ [-2a+b-c]

z=$\frac{1}{2}$ [2a+b+c]

any vector can be written as linear combinations of above equations.

Let $\bar{u} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$

$$x = \frac{1}{3}[-a + 3b - c]$$

$= \frac{1}{3}[-4 + 15 - 6]$

X= 5/3

y= $\frac{1}{2}$ [-2a+b-c] = $\frac{1}{2}[-8 + 5 - 6] = \frac{-9}{2}$

z= $\frac{1}{2}$ [2a+b+c] = $\frac{1}{2}$ [2(4)+5+6]= $\frac{19}{2}$

consider$x\bar{v}_1 + y\bar{v}_2 + z\bar{v}_3$

$$= \frac{5}{3}\begin{bmatrix} 3 \\ 0 \\ -6 \end{bmatrix} + \left(\frac{-9}{2}\right)\begin{bmatrix} -4 \\ 1 \\ 7 \end{bmatrix} + \frac{19}{2}\begin{bmatrix} -2 \\ 1 \\ 5 \end{bmatrix}$$

= 5+18-19

$= 0 - \frac{-9}{2} + \frac{19}{2}$

= -10- $\frac{63}{2}$

$$= \begin{bmatrix} 5 & 18 & -19 \\ 0 & \frac{-9}{2} & \frac{19}{2} \\ -10 & \frac{-63}{2} & \frac{+95}{2} \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

Therefore span $\{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$ = R^3

$\therefore \{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$is basis for R^3

Q.Let $\overline{v_1} = \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}, \overline{v_2} = \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}, \overline{v_3} = \begin{bmatrix} -7 \\ 5 \\ 4 \end{bmatrix}$

Sol: Let β = set $\{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$ to show that β is basis we have to check if first of all β is linearly independent or not span β = R^3

Let c_1, c_2, c_3 be scalars.

$$c_1\bar{v}_1 + c_2\bar{v}_2 + c_3\bar{v}_3 = 0$$

$$c_1\begin{bmatrix}2\\-2\\1\end{bmatrix} + c_2\begin{bmatrix}1\\-3\\2\end{bmatrix} + c_3\begin{bmatrix}-7\\5\\4\end{bmatrix} = 0$$

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

$$R_2 \longrightarrow R_2 + R_1$$

$$R_3 \longrightarrow R_3 - R_1$$

$$\begin{bmatrix}2 & 1 & -7 & 0\\0 & -2 & -2 & 0\\1 & 3 & 15 & 0\end{bmatrix}$$

$R_3 \longrightarrow 2R_3 + 3R_2\,, R_1 \longrightarrow R_1 + R_2$

$$\begin{bmatrix}4 & 0 & -16 & 0\\0 & -2 & -2 & 0\\0 & 0 & 24 & 0\end{bmatrix}$$

$$R_1 \longrightarrow {}^{R_1}/_{4}, R_3 \longrightarrow {}^{R_3}/_{24}$$

$$\begin{bmatrix}1 & 0 & -4 & 0\\0 & -2 & -2 & 0\\0 & 0 & 1 & 0\end{bmatrix}$$

$R_1 \longrightarrow R_1 + 4R_3, R_2 \longrightarrow R_2 + 2R_3,$

$$\begin{bmatrix}1 & 0 & 0 & 0\\0 & -2 & 0 & 0\\0 & 0 & 1 & 0\end{bmatrix}$$

$$R_2 \longrightarrow {}^{R_2}/_{-2}$$

$$\begin{bmatrix}1 & 0 & 0 & 0\\0 & 1 & 0 & 0\\0 & 0 & 1 & 0\end{bmatrix}$$

No. of pivot columns = 3 and no. of variables= 3 therefore the set is linearly independent.

To verify that span$\{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$ = R^3

Let $\bar{u} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ consider the equation $x\bar{v}_1 + y\bar{v}_2 + z\bar{v}_3 = \bar{u}$

$$\begin{bmatrix} 2 & 1 & -7 & a \\ -2 & -3 & 5 & b \\ 1 & 2 & 4 & c \end{bmatrix}$$

$$R_1 \longrightarrow R_3$$

$$\begin{bmatrix} 1 & 2 & 4 & c \\ -2 & -3 & 5 & b \\ 2 & 1 & -7 & a \end{bmatrix}$$

$R_2 \longrightarrow R_2 + 2R_1, R_3 \longrightarrow R_3 - 2R_1$

$$\begin{bmatrix} 1 & 2 & 4 & c \\ 0 & 1 & 13 & b + 2c \\ 0 & 0 & -15 & a - 2c \end{bmatrix}$$

$$R_3 \longrightarrow R_3 + 3R_2$$

$$\begin{bmatrix} 1 & 2 & 4 & c \\ 0 & 1 & 13 & b + 2c \\ 0 & 0 & 24 & a - 2c + 3b + 6c \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & c \\ 0 & 1 & 1 & b + 2c \\ 0 & 0 & 24 & a + 3b + 4c \end{bmatrix}$$

$$R_3 \longrightarrow {}^{R_3}/_{24}$$

$$\begin{bmatrix} 1 & 2 & 4 & c \\ 0 & 1 & 13 & b + 2c \\ 0 & 0 & 1 & \frac{1}{24}[a + 3b + 4c] \end{bmatrix}$$

$$R_1 \longrightarrow R_1 - 2R_2$$

$$\begin{bmatrix} 1 & 0 & -22 & -2b-3c \\ 0 & 1 & 13 & b+2c \\ 0 & 0 & 1 & \frac{1}{24}[a+3b+4c] \end{bmatrix}$$

$$R_1 \longrightarrow R_1 + 22R_3$$
$$R_2 \longrightarrow R_2 - 13R_3$$

$$\begin{bmatrix} 1 & 0 & 0 & \frac{22}{24}(a+3b+4c)+(-2b-3c) \\ 0 & 1 & 0 & b+2c\frac{-13}{24}(a+3b+4c) \\ 0 & 0 & 1 & \frac{1}{24}[a+3b+4c] \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & \frac{11}{12}(a+3b+4c)+(-2b-3c) \\ 0 & 1 & 0 & b+2c\frac{-13}{24}(a+3b+4c) \\ 0 & 0 & 1 & \frac{1}{24}[a+3b+4c] \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & \frac{11a+9b+8c}{12} \\ 0 & 1 & 0 & \frac{1}{24}(-13a-15b-4c) \\ 0 & 0 & 1 & \frac{1}{24}[a+3b+4c] \end{bmatrix}$$

Verification:

$$x = \frac{11a+9b+8c}{12}$$
$$y = \frac{13a+15b+4c}{-24}$$
$$z = \frac{1}{24}[a+3b+4c]$$

Let $\begin{bmatrix}1\\2\\3\end{bmatrix}$ be any vector

i.e., a=1, b=2, c=3

$x = \frac{11a+9b+8c}{12} = \frac{11(1)+9(2)+8(3)}{12} = \frac{11+18+24}{12} = \frac{53}{12}$

$y = \frac{13a+15b+4c}{-24} = \frac{13(1)+15(2)+4(3)}{-24} = \frac{13+30+12}{-24} = \frac{-55}{24}$

$z = \frac{1}{24}[1 + 3(2) + 4(3)] = \frac{19}{24}$

Consider $x\bar{v}_1 + y\bar{v}_2 + z\bar{v}_3$

$$\frac{53}{12}\begin{bmatrix}2\\-2\\1\end{bmatrix} + \frac{-55}{24}\begin{bmatrix}1\\-3\\2\end{bmatrix} + \frac{19}{24}\begin{bmatrix}-7\\5\\4\end{bmatrix}$$

$$\begin{bmatrix} \frac{53(2)}{12} & \frac{-55}{24} & \frac{19(-7)}{24} \\ \frac{53(-2)}{12} & \frac{-55(-3)}{24} & \frac{19(5)}{24} \\ \frac{53(1)}{12} & \frac{-55(2)}{24} & \frac{19\times 4}{24} \end{bmatrix}$$

$$\begin{bmatrix} \frac{212-55-133}{24} \\ \frac{-212+165+95}{24} \\ \frac{306-110+76}{24} \end{bmatrix} = \begin{bmatrix} \frac{224}{24} \\ \frac{48}{24} \\ \frac{22}{24} \end{bmatrix} = \begin{bmatrix}1\\2\\3\end{bmatrix}$$

Span $\{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$ = R^3

$\{\overline{v_1}, \overline{v_2}, \overline{v_3}\}$is basis for R^3

<u>Co-ordinate systems</u>

The Unique representation theorem

Let $\beta = \{\overline{b_1}, \overline{b_2}, \dots\dots\dots\dots\dots\dots\dots \overline{b_n}\}$ be a basis for a vector space v Then for each $\bar{x}$ in V. There exists unique set of scalar $C_1, C_2, \dots\dots\dots\dots\dots\dots\dots\dots\dots, C_n$ such that

$\bar{x} = C_1\bar{b}_1 + C_2\bar{b}_2 + C_n\bar{b}_n$ ----(1)

Proof : Given V is a vector space

$\beta = \bar{b}_1, \bar{b}_2, \dots\dots\dots\dots\dots\dots\dots, \bar{b}_n$ is a basis for V

$\Rightarrow \bar{x} = C_1\bar{b}_1 + C_2\bar{b}_2 + \cdots\dots\dots\dots\dots\dots\dots + C_n\bar{b}_n$----------(1)

Suppose that $\Rightarrow \bar{x} = a_1\bar{b}_1 + a_2\bar{b}_2 + \cdots \ldots\ldots\ldots\ldots\ldots + a_n\bar{b}_n$----------(2)

(1)-(2)=$\bar{x} - \bar{x} = C_1\bar{b}_1 + C_2\bar{b}_2 + \cdots \ldots\ldots\ldots\ldots\ldots + C_n\bar{b}_n - a_1\bar{b}_1 - a_2\bar{b}_2 +$

$\cdots \ldots\ldots\ldots\ldots\ldots - a_n\bar{b}_n$

$\bar{O} = \bar{b}_1(C_1 - a_1) + \bar{b}_2(C_2 - a_2) + \ldots\ldots\ldots\ldots + \bar{b}_n(C_n - a_n)$

$\therefore \beta$ is I all coefficients are equal to zero(0)

$\Rightarrow C_1 - a_1 = 0\,, C_2 - a_2 = 0 \ldots\ldots\ldots\ldots C_n - a_n = 0$

$\Rightarrow C_1 = a_1\,, C_2 = a_2 \ldots\ldots\ldots\ldots C_n = a_n$

$\therefore$ The scalars are unique

Hence the theorem.

Definition : Let $\beta = \{\bar{b}_1, \bar{b}_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots, \bar{b}_n\}$ be a basis for v and $\bar{x} \in v$.

If $\bar{x} = C_1\bar{b}_1 + C_2\bar{b}_2 + \cdots \ldots\ldots\ldots\ldots\ldots\ldots + C_n\bar{b}_n$

Then the weight or scalar $C_1, C_2\,, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, C_n$ are called the coordinates of $\bar{x}$ by $\overline{[x]}_\beta$

$$\overline{[x]}_\beta = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{bmatrix}$$

The mapping $\bar{x} \rightarrow \overline{[x]}_\beta$ is called the co-ordinate mapping (determined by)

1) Consider a Basis $\beta = \{\bar{b}_1, \bar{b}_2\}$ for R^2 where $\bar{b}_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $b_2 \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ suppose $\bar{x}$ in R^2 has the coordinate vectors $\overline{[x]}_\beta = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\overline{[x]}_\beta = \begin{bmatrix} -2 \\ 3 \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$

$$\therefore \bar{x} = C_1\bar{b}_1 + C_2\bar{b}_2 = 2\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

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

2) Let $\beta = \left\{ \begin{bmatrix} 1 \\ -4 \\ 3 \end{bmatrix}, \begin{bmatrix} 5 \\ 2 \\ -2 \end{bmatrix}, \begin{bmatrix} 4 \\ -7 \\ 0 \end{bmatrix} \right\}$ be a basis and $[\bar{x}]_\beta = \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix}$ Find $\bar{x}$

Sol: Let $\bar{b}_1 = \begin{bmatrix} 1 \\ -4 \\ 3 \end{bmatrix}, \bar{b}_2 = \begin{bmatrix} 5 \\ 2 \\ -2 \end{bmatrix}, \bar{b}_3 = \begin{bmatrix} 4 \\ -7 \\ 0 \end{bmatrix}$ Given $[\bar{x}]_\beta = \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix}$

$$\bar{x} = C_1\bar{b}_1 + C_2\bar{b}_2 + C_3\bar{b}_3 = 3\begin{bmatrix} 1 \\ -4 \\ 3 \end{bmatrix} + 0\begin{bmatrix} 5 \\ 2 \\ -2 \end{bmatrix} + (-1)\begin{bmatrix} 4 \\ -7 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 3 \\ -12 \\ 9 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -4 \\ 7 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 & -4 \\ -12 & +7 \\ 9 & \end{bmatrix} = \begin{bmatrix} -1 \\ -5 \\ 9 \end{bmatrix}$$

*****

Unit-II

Probability- Basic terminology, Three types of Probability rules, Statistical independence, Statistical dependency, Bayes theorem.

Probability distributions - random variables, expected values, Binomial distribution, Poisson distribution, Normal distribution, choosing correct distribution.

INTRODUCTION

Probability is a part of our everyday lives. We live in a world in which we are unable the predict the future with complete certainty. Our need to cope with uncertainty leads us to the study and use of probability theory. Probability refers to the chance of occurring or non-occuring of an event. The theory of probability has its origin in the games of chance related to the gambling. It was in the first half of the sixteenth century that serious thought was given to the problem of probability by eminent mathematicians.

PROBABILITY IN EVERYDAY LIFE

Many unpredictable or probabilistic phenomenon where the results cannot be predicted with certainly are frequently observed in economics business and social sciences or even in our day-to-day life. For example:

(i) It toss of fair coin, we are not sure as to whether we shall get a head or a tail.

(ii) A manufacturer cannot ascertain the future demand of his product with certainty.

(iii) In throwing of a die, we are not sure about whether we will get which number 1, 2, 3, 4, 5 or 6.

(iv) In case of a baby to be born the sex cannot be predicted with certainty.

A numerical measure of uncertainty is provided by a very important branch of statistics called the "Theory of Probability". In the words of Prof. Ya-Lin-Chou "Statistics is the science of decision making with calculated risks in the face of uncertainty".

Today the subject has been developed to a great extent and there is not even a single discipline in social, physical or natural sciences where the probability theory is not used. It is extensively used in the quantitative analysis of business and economic problems. It is an essential tool in statistical inference and forms the basis of the 'Decision Theory', viz, decision making under conditions of uncertainty with calculated risk.

BASIC TERMINOLOGY

Random Experiment : An experiment is called a random experiment when it is conducted repeatedly under homogeneous conditions, the result is not unique but may be one of the possible outcomes.

1. Tossing a fair coin is an experiment. Whether the coin will fall head up or tail up is unpredictable.
2. Rolling an unbiased die is an experiment. How many dots it will actually through up is unpredictable and is subject to chance.
3. Drawing a card from a pack of cards.
4. Drawing balls from a bag containing a given number of white and red balls.

Exhaustive Cases. The total number of possible outcomes of a random experiment is called the collectively exhaustive cases of the experiment. For example, in tossing of a single coin, exhaustive number of cases are 2, in throw of a die exhaustive number of cases are 6 and in case of a throw of two dice, exhaustive number of cases are $6^2 = 36$

Equally Likely Cases. The outcomes are said to be equally likely probable if done of them is expected to occur in preference to other. Thus, in tossing of a coin, all the outcomes of head or tail are equally likely if the coin is not biases and if it is the throwing of an unbiased die all the possible outcomes, 1, 2, 3, 4, 5, 6 are equally likely.

An Event. In the theory of probability, the term 'event' is used to denote a phenomenon which occurs with every realisation of a set of conditions. A particular event may be 'simple' or 'compound'.

Event is called simple if it corresponds to a single possible outcome of the experiment, otherwise it is known as a compound or composite event. Thus, in tossing of a single die the event of getting '6' is a simple event but the event getting an even number is a composite event.

(i) In a toss of two coins, the number of cases favourable to the event can be said 'exactly one head' which are 2, viz, HT, TH.
(ii) In drawing a card from a pack of cards, the cases favourable to getting a spade are 13 and 'getting an ace of spade' is only 1.

Mutually Exclusive Events. Two or more events are said to be mutually exclusive if the happening of any one of them excludes the happening of all others in the same experiment. For examples, in toss of a coin the events 'head' and 'tail' are mutually exclusive because if head comes, we can't get tail and if tail comes we can't get head. Similarly in the throw of a die, the six faces numbered 1, 2,3,4, 5, and 6 are mutually exclusive. Thus events are said to be mutually exclusive if no two or more of them can happen simultaneously.

Independent Events. Two or more events are considered to be independent if the occurrence of one event in no way affects the occurrence of the other. The

question of dependence or Independence of events is relevant when experiments are consecutive and not simultaneous. For example, in the tossing of a coin a trail is not affected by the result of the previous left. The events, therefore, are independent.

The following are three approaches used in probability concepts.

(i) Classical approach of Probability

(ii) Statistical or Empirical Probability

(iii) Axiomatic Probability or modern approach of Probability.

CLASSICAL APPROACH OF PROBABILITY

In random experiment there are 'n' exhaustive, mutually exclusive and equally likely outcomes, out of which 'm' are favourable to the happening of an event A, then the probability of occurrence of A, usually denoted by P(A) and is given by

$$P(A) = \frac{Favourable\ number\ of\ outcomes\ to\ A}{Exhaustive\ number\ of\ outcomes}$$

$$= \frac{m}{n}$$

Remarks: 1.The favourable outcomes of the complimentary event $\bar{A}$ are n-m then the probability of non-happening of the event A is given by

$$P(\overline{A}) = \frac{Favourable\ outcomes\ to\ \bar{A}}{Exhaustive\ number\ of\ outcomes}$$

$$= \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - P(A)$$

That implies $P(A) + P(\overline{A}) = 1$

2. Since m and n are positive integers, $P(A) \geq 0$ and the probability of an event is always lies between 0 and 1 ie., $0 \leq P(A) \leq 1$., for any event A.

LIMITATIONS:

The classical probability has fails in the following situations

(i) If n, the exhaustive number of outcomes of the random experiment is infinite.

(ii) If the various outcomes of the random experiment are not equally likely.

(iii) If the actual value of n is unknown.

To overcome the above drawbacks, Statistical or Empirical probability concept is defined.

STATISTICAL OR EMPIRICAL PROBABILITY

Definition : If an experiment is performed repeatedly under essentially homogeneous and identical conditions, then limiting value of the ratio of the number of times occurs to the number of trails, as the number of trails becomes infinitely large, is called the probability of happening of an event, it being assumed that the limit is finite and unique.

Suppose that an event A occurs m times in N repetitions of a random experiment. Then the ratio m/N gives the relative frequency of the event A and it will not vary one trail to another trail. In the limiting case when N becomes sufficiently large, it more or less settles to a number which is called the probability of A.

$$P(A) = \lim_{N\to\infty}(m/N)$$

MODERN APPROACH TO PROBABILITY

The modern concept to probability combines both the objective and subjective concepts of probability. The Russian mathematician Kolmogorov introduced this new approach through the use of the theory of sets. All basic ideas like equally likely events, the favourable outcomes, the character of collectively exhaustive, mutually exclusive event, dependence and independence of events, etc. ., were explained by him through the theory and the operations on sets.

This theory introduces 'probability' simply as a number associated with each event. It is based on certain axioms which express the rules for operating with such numbers. This means that the probability of events must only satisfy these axioms. The advantage of axiomatic theory is that it narrates all situations irrespective of whether the possible outcomes of a random experiment are 'equally likely' or not. It may be noted that classical theory can also be derived from the axiomatic theory as a special case. Now, we define some basic terms and concepts.

1) Sample Space ($\Omega\ or\ S$). The set of points representing all possible outcomes of an experiment is called the sample space and it is denoted by $\Omega\ or\ S$. A particular outcome, i.e., an element in S, is called a sample point.

Thus, if $e_1, e_2, , , , , , , e_n$ are the possible outcomes of a random experiment, the set is said to be the 'sample space' of the experiment. The number of sample points in sample space is generally denoted by $n(S)$.

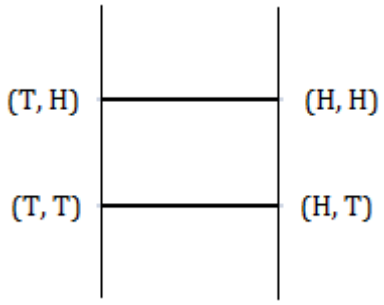
Illustration 1. Consider a simple experiment of tossing a fair coin. In the experiment there are two possible outcomes (i) 'a head up' and (ii) 'a tail up'. Thus the sample space consists of only two sample points. It is convenient to represent the possible outcomes by points on a line 'a head up' is shown by the point 1 and 'a tail up' by the point 0 on the line.

T	H
0	1

Sample space for tossing a coin

Hence the sample space is $S = [0, 1]\ and\ n(S) = 2$.

1. The experiment of tossing two fair coins simultaneously has possible outcomes, namely, HH, HT, TH and TT. It is convenient to represent these outcomes by points $(1, 1), (1, 0), (0, 1) and\ (0, 0)$ in the xy plane.



Sample space of tossing two coins

The sample space may be denoted by $S = [0,1,1,2]$, where each digit is the number of 'head up' in the experiment.

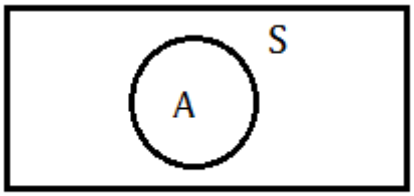
The event sets are denoted by capital letters A, B, C, or E_1, E_2, …, etc. The sample points in each set may be denoted by small letters, say a, b, c, or a_1, a_2, a_3, or any other suitable description. The number of sample points in an event set A may be denoted by $n(A)$.

Mutually Exclusive Events. Two events A and B are said to be mutually exclusive it $A \cap B = \varphi \; i.e.$ if Aand B are disjoint sets, i.e. no sample point is common to both the events and hence they cannot occur simultaneously

Complementary Events. The complement of an event A denoted by $\bar{A}$ is defined as the set of point x such that $x\varepsilon S \; and \; x\epsilon A$. Symbolically,

$$A^c = [x \in S, x \notin A]$$

If the rectangular area denotes the sample space S, and the circle inside denotes the event A, then A^c is the shaded area.



A^c is shaded

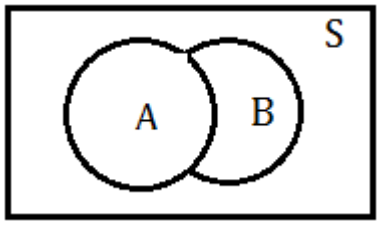
The theoretic presentation of sample space and events enable us to form new events using the set operations.

Unions of Two Events. The union of two events A and B denoted $A \cup B$, is the set of sample points x such that $x \in A \text{ or } x \in B$, symbolically

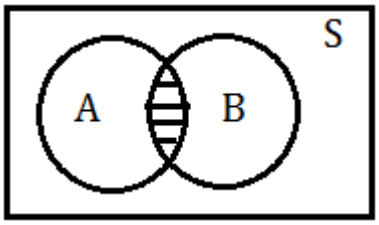
$$A \cup B = [x: x \in A \text{ or } x \in B]$$

Intersection of Two Events. The intersection of two events $A \text{ and } B$ denoted by $A \cap B$, is the set of sample points $x \text{ such that } x \in A \text{ or } x \in B$. Symbolically

$$A \cap B = [x: x \in A \text{ and } x \in B]$$

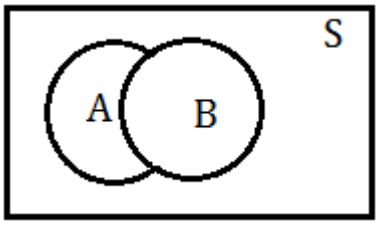


$A \cup B \text{ is shaded}$



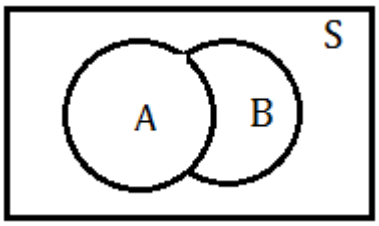
$A \cap B \text{ is shaded}$

We have the following events too-



$(A \sim B) \text{ is shaded}$

Note that $A \sim B = A \cap B$



$(B \sim A) \text{ is shaded}$

Note that $B \sim A = B \cap A$

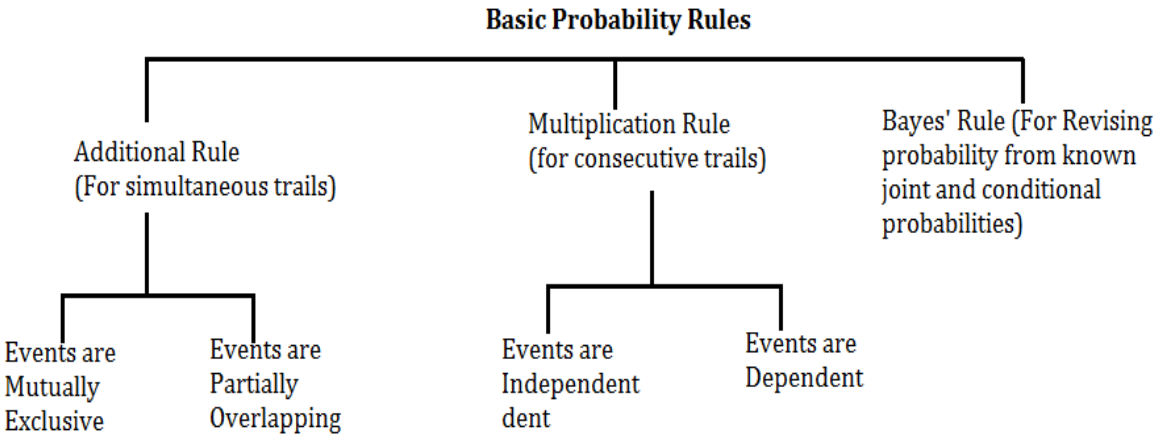
Definition of Probability

Let S be the sample space, let $\{A_i\}$ be the class of events and let f be a real-valued function defined on $\{A_i\}$. Then f is called a probability measure, and P(A) is called the probability of the event A if it satisfies the following axioms:

1) $0 \leq P(A) \leq 1 \text{ for every } A_{i,\, i = 1,2,\ldots,n}$
2) P(S)=1.
3) For every finite or infinite sequence of disjoint events of A_1 , A_2, ..., A_n

   $P(A_1 \cup A_2 \cup \ldots\ldots \cup A_n) = P(A_1) + P(A_2) + \ldots\ldots\ldots + P(A_n)$.

Some important basic rules are as follows



Additional Rule (or Theorem)

If A and B are two events then the probability of occurrence of A or B is given by

P A or B) $= P(AUB) = P(A) + P(B)$ - P $(A \cap B)$

When A and B are mutually exclusive events then

$$P(A \text{ or } B) = P(A) + P(B)$$

Illustration. The probability that a company executive will travel by train in 2/3 and that he will travel by plane is 1/5. The probability of his travelling by train or plane is

P $(A \cup B) = P(Train \text{ or } Plane) = P(Train) + P(Plane)$

$$\frac{2}{3} + \frac{1}{5} = \frac{13}{15}$$

Note: The probability of not travelling by either train or plane.

$$1 - \frac{13}{15} = \frac{2}{15}$$

Illustration 1. A construction company is bidding for two contracts, A and B. The probability that take company will get contract A is $\frac{3}{5}$, the probability that the company will get contract B is $\frac{1}{3}$ and probability that the company will get both the contracts is $\frac{1}{8}$. What is the probability that the company will contract A or B?

Let E_1 be the event that company gets the contract A and E_2 be the event that company gets the contract B. Then by the addition rule, the probability of the Event $'E_1\ or\ E_2'$ (that company, will get contract A or B) is

$$P(E_1\ or\ E_2) = P(E_1) + P(E_2) - P(E_1\ and\ E_2) = \frac{3}{5} + \frac{1}{3} - \frac{1}{8} = \frac{97}{120}$$

CONDITIONAL PROBABILITY

A and B are two events in a sample space , then the conditional probability of A/B (read as A given B) is defined as the probability of A after the occurrence of B and is given by

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0$$

Similarly $P(B/A) = \frac{P(A \cap B)}{P(A)}$, $P(A) \neq 0$

MULTIPLICATION RULE

A and B are two events then the probability of occurring of A and B is given by

$$P(A \cap B) = P(A/B)\ P(B), \quad P(B) \neq 0$$
$$= P(B/A)\ P(A), \quad P(A) \neq 0$$

When A and B events are independent then the probability of occurrence of A and is given by

$$P(A \cap B) = P(A).P(B)$$

In other words,

$$P(A\ and\ B) = P(A).P(B)$$

Illustration. A bag contains 3 while and 4 black balls. One ball is drawn from the bag and then replaced. Another ball is drawn after the replacement. Find the probability that both drawings are of white balls.

Solution. The happening on the first draw certainly has nothing to do with second draw, since the second ball is drawn after the first ball is replaced. Thus, the gtwo draws are independent.

Let

A= the event of the first draw that will have a white ball, and

B= the event of the second draw that will also have a white ball.

Since there are three white balls in the bag of seven balls, in each draw we have $P(A) = P(B) = 3/7$.

Hence $P(A \cap B) = P(A).P(B) = \frac{3}{7} X \frac{3}{7} = \frac{9}{49}$

Illustration. As in previous illustration, assume, that the first ball is not returned to the bag when the second ball is drawn. Find the probability that both balls of the two drawings are white.

Here $P(A) = 3/7$ (since there are three balls in the bag of seven balls).

Event B depends on the occurrence of event 4. If the first ball drawn is white. The probability for the second draw to have a white ball is

$P(B|A) = 2/6$ (Since there are only two white balls left in the bag, of six balls after the first white ball is drawn).

The probability that both balls in the two draws are white is

$P(A\ and\ B) = P(A).P(B \mid A) = 3/7\ .2/6 = 1/7$

INVERSE PROBABILITY

The computation (or revision) of unknown (old) probabilities called priori probabilities (derived subjectively or objectively) in the light of additional information made available by the experiment or past records to derive a set of new probabilities known as posterior probabilities is one of the important applications of the conditional probability shows that it has occurred due to a particular event or reason is called its inverse or posterior probability. These probabilities are computed by Bayes's Rule, named so after the British mathematician. Thomas Bayes who produced it in 1763. The revision of old (given) probabilities in the light of the additional information supplied by the experiment or past records is of extreme help to business and management executive in arriving at valid decisions in the face of uncertainties.

Baye's Theorem (Rule for the Inverse Probability)

If an event B can only occur in conjunction with one of the n mutually exclusive and exhaustive events A_1, A_2, ….. A_n in the sample space S and if B actually happens, then the probability that it was preceded by the particular event A_i (i=1, 2, ……., n) is given by

$$P\left(\frac{A_i}{B}\right) = \frac{P(A_i)P(\frac{B}{A_i})}{\sum_{i=1}^{n} P(A_i)\, P(\frac{B}{A_i})} \quad , \; i = 1, 2, \ldots \ldots, n$$

Example.

Two sets of candidates are competing for the positions on the Board of Directors of a company. The probabilities that the first and second sets will win are 0.6 and 0.4 respectively. If the first set wins, the probability of introducing a new product is 0.8, and the corresponding probability if the second set wins is 0.32. What is the probability that the new product will be introduced?

Solution.

Let the probabilities of the possible events be:

$P(A_1)$ = Probability that the first set wins = 0.6

$P(A_2)$ = Probability that the second set wins = 0.4

P (B) = Probability that a new product is introduced.

$P(B \mid A_1)$ = Probability that a new product is introduced given first set wins = 0.8

$P(B \mid A_2)$ = Probability that a new product is introduced given second set wins = 0.3

Then the rule of additions gives:

P (New Product) = P (First set and new products) + P (Second set and new product)

In terms of symbols:

$P(B) = P(A_1 \cap B) + P(A_2 \cap B) = P(A_1)\,P(B \mid A_1) + P(A_2)P(B \mid A_2)$

$= 0.8 \text{ X } 0.6 + 0.3 \text{ X } 0.4 = 0.60$

RANDOM VARIABLE

A random variable means a real number associated with the outcomes of a random experiment. It can take any one of the various possible values each with a definite probability. For example, In tossing of two coins if X denotes the number of Heads, then X is a random variable which can take any oneof the values 0 (no head), 1 (with single head) and 2 (both are heads) each with probability ¼, 2/4,and ¼ respectively.

Definition: Random variable is a real valued function on the sample space, taking values from $R(-\infty, \infty)$. i.e., Random variable is real valued function from sample space to real numbers. ie., $f : S \rightarrow R$.

If the random variable, X, Takes countable number of values or finite values it is known as discrete random variable, foe example marks of group students in a

test is a discrete random variable. If X takes infinite number of values or uncountable number of values it is called continuous random variable. For Example, The Heights, weights of a group of persons is a continuous random variable.

PROBABILITY DISTRIBUTIONS

We are now discussing inference statistics in which we draw inference on the nature of the populations by the study of a sample. This is the inductive process of drawing an inference. As against this, in the deductive process we estimate what the sample would be like when the population parameters are known.

In this process we have to make use of probability distribution These theoretical probability distribution are like the relative frequency distributions and these total up to one. Further, whereas the frequency distributions were based on actual observations, the theoretical distributions ware based on mathematical models. An important property of these theoretical distributions is that with some known parameters we may know the whole distributions. They have also some well defined properties such as the mean, variance, standard deviation, skewness; which make them particularly useful.

However, a probability distribution may not fully agree with an empirical distribution; but with a large number of trails there is a likelihood that the observed values will approach the theoretical values. The theoretical distribution serves as a model, for testing the 'goodness of fit' of the empirical distribution. It is because of this that they are also called the normal curves of error: with the help of these we can know the extent of error in the estimate.

The following are some important theoretical distributions used for drawing statistical inference. The probability distributions are two types.

a) a)Discrete probability distributions : Binomial distribution, Poisson Distribution, Geometric Distribution etc.,
b) Continuous probability distributions : Normal distribution, Uniform distribution, Gamma distribution etc.,

BINOMIAL DISTRIBUTION

The Binomial distribution is one of the most widely used probability distribution of discrete random variable. This distribution is also known as Bernoulli distribution since it was introduced by a Swiss mathematician, J Bernoulli (1654 – 1705). The Binomial distribution describes the distribution of probabilities where there are only two possible outcomes for each trail. of an experiment, those are called success and failure. Let E be an event of Bernoulli trail. Let the probability of success is p and probability of failure is q then p + q =1.

Definition: In n independent trails probability of getting 'r' successes and 'n - r' failures is given by

$P(X = r) = {}^{n}C_r\, p^r q^{n-r}$ r = 0,1,2, ...,n

0, otherwise

Where p = probability of success
q = probability of failure
n = number of independent trails
r = number of success
and p + q = 1

Here n and p are the parameters of Binomial distribution.
Also Mean = np and Variance = npq

Eg 1. A coin is tossed 9 times. Find the probability of getting 5 Heads.

Solution: $B(x,n,p) = {}^{n}C_x p^x (1-p)^{n-x}$

Here n = number of trials =9

X =5, p = probability of getting head = $\frac{1}{2}$

$\therefore$ Required probability $= {}^{9}C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^{9-5} = {}^{9}C_5 \left(\frac{1}{2}\right)^9 = \frac{63}{256}$

Eg 2. A die is thrown 8 times. If getting a 2 or 4 is a success. Find the probability of

(i) 4 success
(ii) P ($X \leq 3$)
(iii) P ($X \geq 2$)

Solution: $B(x, n, p) = {}^nC_x P^x(1-p)^{n-x}$

N = 8 = number of trials = 8

P = probability of getting 2 or 4 = $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

i) x = 4

$p(X=4) = {}^8C_4\left(\frac{1}{3}\right)^4\left(1-\frac{1}{3}\right)^{8-4} = {}^8C_4\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^4 = \frac{1120}{6561}$

ii) $p(X \leq 3) = p(X=0) + p(X=1) + p(X=2) + p(X=3)$

$= {}^8C_0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^8 + {}^8C_1\left(\frac{1}{3}\right).\left(\frac{2}{3}\right)^7 + {}^8C_2\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^6 + {}^8C_3\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^5 = \frac{4864}{6561}$

ii) $p(X \geq 2) = 1 - p(X=0) - p(X=1)$

$= 1\left(\frac{2}{3}\right)^8 - \frac{2^7}{3^8} = \frac{5281}{6561}$

Eg. 3. Among the items produced in a factory 5% are defective. Find the probability that a sample of 8 contains

i) exactly 2 defective items
ii) greater than or equal to 7 defective items
iii) at least one defective items.

Solution: $B(x,n,p) = {}^nC_x p^x (1-p)^{n-x}$

Here n = 8

P = probability of defective item = $\frac{5}{100} = \frac{1}{20}$

i) x = 2

$= {}^8C_2\left(\frac{1}{20}\right)^2\left(\frac{19}{20}\right)^6 = \frac{7}{100}\left(\frac{19}{20}\right)^6$

ii) $p(X \geq 7) = P(X = 7) + P(X = 8)$

$= 8C_2\left(\frac{1}{20}\right)^7 .\frac{19}{20} + \left(\frac{1}{20}\right)^8 = \frac{153}{(20)^8}$

iii) p (atleast one defective item) = $P(X \geq 1)$

$= 1\text{-}P(X=0) = 1\text{-}\left(\frac{19}{20}\right)^8 = \frac{1-(19)^8}{(20)^8}$

Example: Find the binomial distribution whose mean is 9, the variance being 2.25

From the description of the problem we may write up = 9 and $npq = 2.25$.

Now $q = \frac{(npq)}{(np)} = \frac{2.25}{9} = 0.25 \ \ and \ \ consequenthy \ p = 1 - q = 0.75;$

And $q = \frac{np}{p} = \frac{9}{0.75} = 12.$

The required binomial distribution is $(q + p)^n = (0.25 + 0.75)^{12}$

POISSON DISTRIBUTION. In binomial distribution, there is a sample of a definite size and we can count the number of times a certain event is observed. There are problems however, in which the number of times an event occurs can be counted without there being any sense is asking how many times the event did not occur. For example, the goals served in a football match or accident in a factory, to all such cases the binomial distribution is in applicable precisely because we do not know the value of n in the fundamental expression $(p + q)^n$.

To deal with events of the type we make use of the Poisson distribution. It is necessary to know a mathematical constant denoted by the letter e. it arises in the study of the natural law of growth (the exponential law) and has the value:

$$e = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ \ldots\ldots\ldots\ldots\ldots\ldots$$

=1+1+0.5+0.1667+0.04167+0.0833+0.00139+0.00020+0.00002+. . . . . which gives e=2.7183 correct to four decimal places.

The following are the properties of a Poisson Distribution:

Mean = $np \ or \ m \ or \ z$

Variance = $np \ or \ m \ or \ z$

Standard Deviation $= \sqrt{np} \ or \ \sqrt{m} \ or \sqrt{z}$

POISSON DISTRIBUTION

Definition: Poisson distribution is a limiting case of Binomial distribution under the following conditions:

(i) The variable is discrete

(ii) A dichotomy is evolved

(iii) Statistical independence is assumed

(iv) n, number of independent trails are very large.

(v) p or q is very small and close to zero, or unity. If p is close to zero, the distribution will be J-shaped and unimodal.

The probability function of Poisson distribution is given by

$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!}$, r = 0,1,2,.....

where λ = the arithmetic mean,

r = number of successes

Here λ is the parameter of the distribution and the special feature of poisson distribution is mean = λ and variance = λ

And e=2.71828

If the probability that an individual suffers a bad reaction from a certain injection is .003. Find the probability that out of 1000 individuals i) exactly 3 ii) mote than or equal to 2 individuals iii) None suffers from a bad reaction.

μ =np; n=1000; p =003; μ = 1000 x .003 =3

i) x =3

Probability that 3 suffers among 1000 is p (3, 3)

$p(x.\mu) = \frac{e^{-\mu}.\mu^x}{x!}$

$p(3, 3) = \frac{e^{-3}.(3)^3}{3!} = \frac{9}{2}e^{-2} = \frac{9}{2}(.0498) = .2241$

ii) $p(x \geq 2) = 1-(p(x=0) + p(x=1)$

$= 1-e^{-3} - e^{-3}.3 = 1 - e^3(4)$

$= 1 - 4x (.0498)$

$= 1-.1992$

$= .8008$

iii) $p(x=0) = p(0, 3) = e^{-3} = .0498$

2. 2% of the items of a factory are defective. The items are packed in boxes. What is the probability that there will be

(i) 2 defective items

(ii) atleast three defective items

(iii) 2 < defective items <5) in a box of 100 items.

Solution: $\mu = np; n = 100,$ p = probability of defective items.

$= \frac{2}{100} = .02$

$\mu = .02 \times 100 = 2$

i) x = 2

$p(2, 2) = e^{-2}.\frac{(2)^2}{2!} = 2e^{-2} = 2 \times (.136) = .272$

ii) $p(x \geq 3) = 1-p(x=0)-p(x=1) - p(x=2)$

$= -e^{-2}-2e^{-2}-e^{-2}\frac{(2)^2}{2!} = 1-5e^{-2} = 1-5 \times .136 = .320$

iii) $p(2<x<5) = p(x=3)+p(x=4)$

$= \frac{e^{-2}.(2)^3}{3!} + \frac{e^{-2}.2^4}{4!} = 2e^{-2} = .272$

3. The probabilities of a Poisson variate taking the values 1 and 2 are equal. Calculate the probabilities of the variate taking the values 0 and 3.

Solution: $p(x, \mu) = \frac{e^{-\mu}.\mu^x}{x!}$

$P(1, \mu) = p(2, \mu)$

$e^{-\mu}.\mu = \frac{e^{-\mu}.\mu^2}{2!} \therefore \mu = 2$

$p(x=0) = p(0, 2) = e^{-2} = .136$

$p(x=3) = p(3,2) = \frac{e^{-2}.2^3}{3!} = \frac{4e^{-2}}{3} = .181$

NORMAL DISTRIBUTION: The Binomial and Poisson distributions discrete distributions and enable us deal with the occurrence of distinct events, events such as the number of defective items in a sample of a given size or the number of accidents occurring in a factory, the normal distribution is a mathematics distribution for dealing with quantities whose magnitude vary continuously.

Normal Distribution: The probability function of Normal distribution is denoted by f(x) and is given by

$f(x) = \frac{1}{\sigma\sqrt{2\pi}}$, $-\infty < X < \infty$

$-\infty < \mu < \infty$ and $\sigma^2 > 0$

Here μ (mean) and σ^2(variance) are called the parameters of the Normal distribution.

Chief characteristics of normal distribution

1. The mean, median and mode for normal distribution are identical.
2. The curve is smooth, regular, bell – shaped and symmetrical about the line

   $x = \mu$ since the curve $y = f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ does not have the odd powers of $\left(\frac{x-\mu}{\sigma}\right)$.

3. The ordinate of the curve decreases rapidly as |x| increases, the maximum ordinate at x =μ is given by $y_{max} = \frac{1}{\sqrt{2\pi\sigma}}$ and the curve is unimodal.

4. The curve extends from -∝ to ∝.

5. As σ becomes larger, the ordinate y decreases i.e., the curve spreads out more but flattered at the top. On the other hand when σ becomes smaller y increases and the curve becomes more peaked.

6. Total area under the curve above x axis from -∝ to ∝ is unity.

   For a normal distribution with mean μ, and s.d . σ, the total are under under the normal curve is I and

   (i) about 68% of the area fall between $\mu - \sigma$ and $\mu + \sigma$
   (ii) about 95.5% of the area fall between $\mu - \sigma$ and $\mu + 2\sigma$
   (iii) about 99.7% of the area fall between $\mu - \sigma$ and $\mu + 2\sigma$

7. The area bounded by the curve with x-axis and any two ordinates equals to the probability for the interval marked as x axis by the two ordinates.

   Importance of Normal Distribution

   Normal distribution plays a very important role in statistical theory.

   1. Most of the distribution occurring in practice example Binomial, Poisson, Hyper geometric distributions etc., can be approximated by normal distribution moreover, many of the sampling distributions students t, x^2, distributions etc tends to normality for the large samples.
   2. Many of the distributions of sample statistic, the distributions of sample mean, sample variance etc., tends to normality for large samples and as such they can be studied with the help of the normal curve.
   3. The entire theory of small sample tests based on the fundamental assumption that the parent populations from which the samples have been drawn follow normal distribution.
   4. Theory of normal curves can be applied to the graduation of the curves which are not normal.
   5. Normal distribution finds large application in statistical quality control in industry for setting control limits.

   Eg1. If X is a normal variate with mean 30 and standard deviation 5. Find the probabilities that.

   i) 26≤x≤40 and

   ii) x ≥ 45

i) $z = \frac{x-\mu}{\sigma}$

μ = mean = 30

σ = standard deviation = 5

$Z_1 = \frac{x_1-\mu}{\sigma}$, x_1 =26

$Z_1 = \frac{26-30}{5} = -.8$

$Z_2 = \frac{x_2-\mu}{\sigma}$, x_2 =40

$Z_2 = \frac{40-30}{5} = 2$

P(26≥x≤40) =p (-.8≤z≤2)

P(-.8≤z≤2)

= p(-.8≤z≤0) +p(0≤z≤2)

= p(0≤z≤.8) +p(0≤z≤2) by symmetry

= .2881 +.4772 =.7653 (from tables)

ii) $X \geq 45$

$z = \frac{45-30}{5} = 3$

p(X≥45) = p(z≥3)

= 0.5 - Area (z=3)

= 0.5 – 0.4987=.0013

∴ p(X≥45) = .0013

These limiting cases approach what is popularly known as the normal distribution. Unlike both the binomial and the Poisson distribution it is continuous, as shown below:

A random variable x is said to have a normal distribution with parameter $\mu(= x)$ and σ if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-(x-\mu)^2/2\sigma^2} \quad -\alpha < x < \alpha$$

Such a random variable is denoted by N(μ, σ^2). These two parameters describe the distribution completely. It is to be noted that the probability density has the maximum value at the mean , and it decreases gradually on either side.

It is possible to convert any distribution to the standardised form. This is because the normal distribution has the same shape whatever is parameters (mean and standard deviation) may be.

The purpose of standardization of the normal distribution is that we can make use of the tables of the area of the standard curve (representing probability)

$f(z) = \frac{1}{2\Pi} e^{-z^2/2}$ for the various points along the X-axis

The standard table is given in any text. Similar tables of the ordinates of the standard curve are also given in any text. The use of these tables should be clear from the following example:-

Example: -

Suppose that X is a continuous variable with a normal distribution and has a mean of 12 and standard deviation equal to 2. What is the probability that the value of X selected at random lies between 11 and 14?

The first step is to change from the X scale to Z scale

$$For\ X = 11, z = \frac{X - \bar{x}}{\sigma} = \frac{11 - 12}{2} = -\frac{1}{2} = 0.50$$

$$For\ X = 14, z = \frac{X - \bar{x}}{\sigma} = \frac{14 - 12}{2} = 1$$

We are, therefore, interested in the probability that z lies between, 0.50 and 1

Thus $P(11 \leq X \leq 14) = P(-0.50 \leq z \leq 1) =$

$P(0 \leq z < 0.50) + P(0 \leq z \leq 1) = 0.1915 + 0.3413 = 0.5328$

Example: Assume that the mean height of the soldiers to be 68.22 with a variance of 10.8. How many soldiers in a regiment of 1,000 would you expect to be over 6' tall.

$$\sigma = \sqrt{10.8} = 3.286$$

X-μ = 72-68.22 = 3.78

Resorting to the normal probability curve

$$z = \frac{X - X}{\sigma} = \frac{3.78}{3.286} = 1.15$$

Area between X=0 and X=1.15 for the normal probability curve from the tables=0.3746. Therefore area to the right of the ordinate at

1.15 = 0.5 - 0.3746 = 0.1254

This is the required probability. Out of 1,000 soldiers 1000 x 0.12354 =1234 soldiers are expected to be over 6'.

RELATION BETWEEN BINOMIAL AND POISSON DISTRIBUTIONS

In the binomial distribution if n is infinitely large, the probability p of occurrence of events is closed to zero and $q = (1 - p)$ is close to 1. In such cases the binomial distribution is very closely approximated by the Poisson distribution with the mean m=np. This can be ascertained from the properties given earlier by placing $m = np$ and $q = 1$ and $p \approx 0$ (stands for nearly equal).

Since there is a relation between the binomial and normal distributions, it follows that there is also a relation between the Poisson and Bormal distributions. It can in fact be shown that the Poisson distribution approaches a Normal distribution with standardised variable $(x - m)\sqrt{m}$ as m increase indefinitely.

RELATION BETWEEN BINOMIAL AND NORMAL DISTRIBUTIONS

If neither p nor q is very small but n is sufficiently large, the binomial distribution is very closely approximated by the Normal distribution with the mean $m = np$ and $s.d\ .\sigma = \sqrt{npq}$.

Unit-III

Sampling and sampling distributions – Random sampling, Non-Random sampling distributions, operational considerations in sampling.

Estimation – Point Estimates, Interval Estimates, Confidence intervals, calculating interval estimates of the mean and proportions, t-distribution, determinates of sample size in estimation.

SAMPLING THEORY

INTRODUCTION:

POPULATION OR UNIVERSE

The whole from which a sample is drawn is called universe or population. Such a universe or population should be defined precisely and carefully. The listing of all the sampling units in the universe is defined as a sampling frame. When each unit in the population is numbered for identification it would be called a sampling frame. Such a frame helps in identifying any particular item in the population such as electoral list of households, the layout map of a town, etc.

A universe may be finite or infinite. In the former, the size of the universe can be known precisely and a number can be assigned to each unit e.g., accounts receivable in a ledger. While in case of the latter assignment of the number is not possible as in the case of bags of wheat.

A universe may real or hypothetical. In the former case, the universe is known and is in existence at the time of sampling. While in case of the latter, the universe is built up by repeating the event any number of times as in the case of throwing of a dice or tossing of a coin.

Statistical data may be collected by complete enumeration called census inquiry by partial enumeration called sample inquiry. In the former case the information is collected about each and every item comprising the whole (called universe or population is statistical parlance) while in the latter case information is collected about a small number of items which are representative of the whole so as to

form an estimate of the characteristics of the whole . If such a sample is adequate representative of the whole, is properly drawn and interpreted, then it is most likely to represent the conditions of the whole and can be fairly relied upon as if the observation had been based on complete enumeration. To draw a simple example from day to day life, such a method is used and when a examines a few particles of rice from a kettleful to form an idea about the state of ripening of the whole lot. Similarly, I the field of business while the accounting results are based on compilation of business transactions, test checks or test audits cover only a small number of entries to verify the truth of all the entries. Thus, the primary object of sampling is to obtain the maximum of information about population with the minimum effort, and also to set out the limits of accuracy of estimates based on sampling.

“It may be too expensive or too time consuming to attempt either a complete or a nearly complete enumeration in a statistical study. Furthermore, to arrive at valid conclusions, it may not be necessary to enumerate all or nearly all of a population. We may study a sample drawn from the larger population and if that sample is adequately representative of the population, we should be able to arrive at valid conclusions". - P.E. Croxton and D.J. Cowden

Advantages of sampling techniques :

Samples are devices for learning about large masses by observing a few individuals. The following are some of the advantages of making use of the sampling technique:

(i) It reduces the cost because only a few selected items are studied in sampling. The characteristics of the population can be known from this sampling.

(ii) It saves time because the data can be collected and summarised more quickly with a sample than with a complete count. This method is very useful when results are urgently required.

(iii) Sampling method is the only method which can be used in the case of infinite population.

(iv) A sample may actually produce more accurate results than the kind of complete enumeration.

(v) It is difficult to handle a population which usually consists of a large number of units.

Drawbacks of sampling techniques:

However, the sampling techniques have the following drawbacks:

(i) Sometimes the population may be so small that it may be impossible to draw a representative sample from it.

(ii) The result may be false, inaccurate and misleading if the sample has not been drawn properly.

(iii) There may be personal bias and prejudice with regard to the choice of technique and drawing of sampling units.

Essentials of sampling:

A sample should possess the following essentials for valid conclusions of the experimental results:

(i) A sample should have the similar characteristics of the original population from which it has been selected.

(ii) Selected samples from the population should have similar nature. There should not be any difference when compared with the population.

(iii) The number of observations included in a sample should be more to make the results more reliable.

Size of the sample:

The size of the example is the quantity of inspecting units, which are chosen from a populace for examination. The issue that emerges is, the way to conclude the size of the example which ought to be chosen from a populace. Assuming that the

size of the example is little, it may not address the populace and it won't be imaginable to find out the exactness in results. Then again, on the off chance that the size of the example is enormous, more prominent will be the portrayal of the things of the populace in it and it will be very challenging to make due. In this manner, the example size ought to be neither too large nor excessively little. It would be ideal for it to be "ideal". The size of the example relies upon various contemplations which are as per the following:

(i) In a population consisting of perfectly homogeneous units, a small sample may serve the purpose. Where as in a population consisting of heterogeneous units, a large size sample is inevitable for yielding reliable results. For example, the blood of a person in perfectly homogeneous and, hence. a drop of blood taken for investigation gives the true picture of the blood constitution in the body.

(ii) The larger the size of the population, the bigger should be sample size.

(iii) The nature of the study also affects the size of the sample. For an intensive and continuous study, a small sample may be suitable. However, in studies which are not likely to be repeated, it may not be necessary to take a large sample size.

(iv) In case it is necessary to classify data in a large number of classes, a large sized sample should be taken to ensure reliability of results.

(v) The availability of trained personnel, finance and time, and other practical considerations also constitute a big constraint on the sample size.

(vi) It is believed that if the sample size is large, there will be a great degree of accuracy. However, it is not always true. If a sample is selected by experts through scientific method, a small sample can give better results.

(vii) The size of the sample is also influenced by the sampling technique. For example, in Random sampling, greater accuracy in results will be achieved only in a large sample. However, in stratified sampling, even a small size gives better results.

METHODS OF SAMPLING:

Method of sampling: There are many different ways of selecting a sample. The following are some of the important methods employed in sampling. There are two different ways of taking an example; one is called irregular, change or likelihood inspecting and the other non-random examining.

RANDOM SAMPLING METHODS:

Random sampling:

A random sample is one where each item of the population has an equal chance of being included in the sample. A random sample may be taken from an infinite or finite population. When population is infinite, "A random is considered a random sample as long as each observation that has been taken does not affect the probability that any other observation will be selected". In case the sample is taken from a finite population, the probability of each observation of the population included in the sample does not remain unaffected. In such a case, "A random sample is one that is drawn in such a way that every available potential observation in the population has an equal probability of being selected in the sample".

Random sampling is a scientific method of getting a sample from the population. This method is also known as "unrestricted random sampling" device. It refers to that technique of sampling in which each and every item of the population has the same probability of being included in the sample. It completely depends on the element of chance. Several methods have been adopted for random selection of the sample. However, to ensure the randomness of selection, one may adopt either the lottery method or random numbers method.

(a) Lottery method:

This is the most popular and simplest method of selecting a random sample from a finite population. In this method, all items of population are numbered on separate slips of paper of identical size, shape and colour. These slips are folded and mixed up in a box and a blind fold selection is made. For a required sample size, the same number of slips are selected. It indicates that the selection of each item thus depends on chance. This lottery method is common in all area of sciences. For example, suppose we have a population of 50individuals in a field of magnitudes 1 to 50 units and we wish to take sample of 10 individuals randomly from the population of 50. Now we must write the numbers of all the 50 individuals on slips of the same size, shape and colour, mix them up and a blind fold selection of 10 slips is to be made (replacing each slip once it has been drawn out). The numbers corresponding to the slips drawn will constitute the random sample. This method is also called unrestricted random sampling because units are selected from the population without any restriction. If the population is infinite this method is inapplicable.

(b) Random numbers:

For use of such tables, it is necessary to have the total population numbered from 1 to N; it would then be possible to determine the range of numbers to be sampled for obtaining the number of digits required to give very unit in the population a chance of being selected. If the population size consists of less than 100 units, two digit numbers are selected; if it contains less than 1,000 units but more than 100 units, 3 digit numbers are required. Suppose, the size of N is 3928 and sample size (n) is 10. Since N= 3928 is greater than 1,000 but less than 10,000, 4 digit numbers are to be selected from the table of random sampling numbers. The reason is that if 3 digit numbers are selected from 0 to 999, some of the units of population will never be selected. The number may be picked up vertical, horizontally or diagonally. For a proper starting point, any six digit figure may be picked up at random from the table and then used; the first two

digits for page number, next to for row and last for the column. Thereafter the table may be read out in a given order vertically, horizontally or diagonally; leaving out any number falling out of the sample frame until the required number of sample units have been picked up.

Now with the availability of tables of random numbers, the work of drawing random samples has become simple. One can use the table of random numbers from any position either horizontally or vertically. A suitable way to choose a starting point is to put a pencil blindly on some number and start reading from this point.

Merits and demerits of random sampling:

Merits:

(i) This method is more scientific because there are less chances for personal bias in sampling from the population. Every item of the population has equal chance of being selected.

(ii) Sampling error can be measured.

(iii) When the size of the sample increases, it becomes increasingly representative of the population. Therefore, in such situations, judgement sampling cannot be used.

(iv) This method is economical as it saves time, money and labour in investigating a problem.

Demerits:

(i) This method requires a complete list of all the items of the population. However, such up to date lists are not available in many enquiries.

(ii) In situations where the size of the sample is small it will not be a true representative of the population.

(iii) Random sampling will not be possible where only certain data are accessible.

(iv) If the units of the population are spread over a large area, this method cannot be used.

Stratified sampling:

When a population is heterogeneous with respect to variable or characteristic under study, stratified random sampling will yield better results when compared to other methods of sampling. If a population is divided into relatively homogeneous groups or strata and a random sample is drawn from each group or stratum to produce an overall sample, it is known as stratified random sampling. The division of the population into strata or groups is done according to some relevant characteristics. Each stratum is also called sub population. It is a method of sampling for giving representations to all strata of population or society such as selecting a sample from areas, classes, sexes etc. This method gives a more representative sample than random sampling in a given large population. There are two types of stratified random sampling. They are proportional and non-proportional. Proportional stratified sampling is one in which the items are taken from each stratum in the proportion of the units of the stratum to the total population. If the number of items are large in the population, the same sampling will have a higher size and vice versa. On the other hand, in non-proportional stratified sampling, an equal number of units are taken from each stratum irrespective of its size.

However, one should take care of the sample size in the case of heterogeneous population. If it is known that the variability is greater in the population, it is better to take a large sample to achieve.

Merits and demerits of stratified random sampling:

Merits:

(i) It is more representative. For example, in random sampling, though each item in the population has an equal chance of being selected, yet due to chance, important groups remain unrepresented. However, in the case of stratified random sampling, every group is being represented in the sample.

(ii) Stratified random sampling ensures greater accuracy, since variability within each stratum is considerably less than the variability in a random sample.

(iii) The units from the different strata may be selected in such a way that all of them are localised in one geographical area.

(iv) In a non-homogeneous population, it may yield more reliable results.

Demerits:

If proper stratification of the population is not done, the sample will have an effect of bias. If different strata of the population overlap each other, it is difficult to draw a sample which should be a representative one.

Systematic Sampling:

Systematic sampling is a predetermined procedure . In this sample selecting the first observation is the important, the remaining observations are automatically selected. Sometimes we might need the sample of trees from a forest or houses in a city. In such cases, a sampling plan known as systematic random sampling is applied. According to this method, a list of the population is prepared on some basis. For this, we arrange the items in numerical, alphabetical or geographical or any other order. Now the items are serially numbered. The first item is selected at random. For example, if we want to select a sample of 10 trees from 100 trees of a forest by taking every k th tree where 'k' refers to the sampling interval. Symbolically;

$$k = \frac{N}{n}$$

Where k = sampling interval

N = population size

n = sample size

Therefore, k = 100/10 = 10. Now 10 is the sampling interval. Every 10th tree will be taken as a sample, i.e. 10th, 20th, 30th, 40th and so on. Systematic sampling is

relatively a simple technique, and may be more efficient than the simple random sampling, provided that the list of observations are arranged at random.

Merits and demerits of systematic sampling:

Merits:

The systematic random sampling is a relatively simple and convenient method of sample selection. The time and labour involved in systematic sampling are relatively less. Even if the populations are arranged, systematic sampling will yield better results.

Demerits:

The main demerits of this method is that it may not represent the whole population.

Cluster sampling:

When there is an unequal concentration of individual units in the universe, this method is employed whereby certain blocks or clusters of higher concentration are selected for complete inquiry: e.g., all cards from one or more of the ledgers, all transaction of one are more weeks in a year or all accounts beginning with the particular alphabet which show highest concentration. These clusters are used often in multistage sampling latter in this study.

It is also known as sampling stages. In the cluster sampling method, the population is divided into some recognisable subgroups which are called clusters. Now the random sample of these clusters is drawn and all the units belonging to the selected clusters constitute the sample. However, it refers to the sampling procedure which is carried out in several stages. Cluster sampling is widely used for geographical studies of many kind. When the units are spread over a large geographical area, selecting a sampling unit becomes expensive. The area may be divided into convenient subgroups called clusters; select a sample of

clusters and collect the data on all units in each of the selected clusters In this method, the clusters should be of small size and the number of sample units in each cluster should be more or less the same.

The advantage of this method is that it possesses flexibility which is lacking in other methods. Another advantage is the large scale survey where the preparation of the list is difficult, time consuming or expensive but this method is less accurate than any other method of selecting a sample by a single stage process.

NON-RANDOM SAMPLING METHODS

Judgement, purposive or deliberate sampling:

Under this method the selection is often based on certain predetermined criteria. The fixation of criteria and deliberate choice of the sampling units which fit into the criteria bring in the personal element and introduce bias into the system. Specially, in case of auditing, an auditor is often interested in knowing the worst about the population e.g., the cases of undue inventory accumulation, delinquent accounts receivable, etc. The application of such a method would bring in bias of the person; the selection would differ from person to person, guided at times by personal fancy and convenience. Thus, the result would be as good as the judgement of the individual determining the sample.

Merits and demerits of judgement sampling

Merits:

(i) It is a simple method.

(ii) It is used to obtain a more representative sample.

(iii) In case the size of the sample of the population is small, the random selection may miss the important items of the population. In such cases, use of judgement sampling is justified.

(iv) This method is widely used in solving every day business problems and making public policy decisions.

Demerits:

(i) Due to individual bias, the sample may not be a representative one.

(ii) The estimates are not accurate.

(iii) Its results cannot be compared with other sampling studies.

Convenience sampling:

In the convenience sampling selection method, selection of items results in obtaining a chunk of the population. A "chunk" is a convenient slice of a population which is commonly referred to as a sample. However, the results obtained by convenience sampling method hardly be representative of the population. They are generally biased and unsatisfactory. For example, a list of students of a University may be suitable for an enquiry into the matters pertaining to education in the University but not a University education as a whole in the country.

Quota sampling:

Quota sampling: In this method each person engaged in the primary collection of data is assigned a certain quota of investigations. Although certain criteria are prescribed for the selection of respondents, the actual choice of all such respondents or a part of the whole quota, who are by any reason not approachable is left to the investigator who is permitted to substitute others to fulfil the quota assigned to him. This method is often adopted in marketing research studies where it is not possible to stick to it without delay and expenditure. Thus this method also allow some bias to enter into the inquiry. Quota sampling is a type of judgement sampling. This method is most commonly used in non-probability categories. Therefore, personal prejudice and individual

bias are there, although this method involves less money and time as compared to other methods. So the quota sampling is not very popular.

CHOICE OF SAMPLING METHODS

The different methods discussed earlier can be used in different situations, wherever they are appropriate. However, it is difficult to say that a particular method would always be better than the rest. Each method has its own speciality. No one method can be regarded as the best under all circumstances. But a number of factors such as the nature of the problem, the size of the sample, the size of the population, availability of finance, time etc. would influence the selection of a particular method of sampling.

However, it is not necessary that the random sampling is always better than judgement sampling. One should take into consideration the merits and demerits of both the methods. Where the size of the sample is small in relation to the size of the population, judgement sampling would yield better results than the random sampling. If the size of the sample increases, random sampling would be more appropriate. In some cases, stratified sampling may give better results than random sampling as well as judgement sampling. There may be some situations where cluster sampling may give better results. We can conclude this by saying that no individual sample plan can be recommended for individual adoption. The choice of the sampling plan must be decided according to the different factors.

SAMPLING AND NON-SAMPLING ERRORS

It is necessary to understand clearly the role of sampling and non-sampling errors in complete enumeration and sample surveys. Errors that arise due to drawing inferences about the population on the basis of a sample are termed sampling error. According to Patterson, "Sampling error is the difference between the result of the census of the whole population". However, errors that mainly arise at the stages of observation and processing of the data, can appear

both in complete enumeration and sample survey, are called non-sampling errors.

SAMPLING ERRORS:

Sampling errors will arise in sample which is subset of population. The reason is that estimate is based on a part and not on the whole population. Hence, sampling gives rise to certain errors known as sampling errors. Sampling errors are of two types: biased and unbiased.

(i) Biased errors: These errors arise because of bias in selection, estimation etc. For example, deliberate sampling method may be adopted in the place of simple random sampling method where some bias is introduced in the result and, hence, such errors are called biased sampling errors. Such errors are also known as cumulative errors or non-compensating errors because they do not decrease in a large population even after increase in the sample size.

(ii) Unbiased errors: Unbiased errors arise due to chance differences between members of the population included in the sample and members not included in the sample. It is known as random sampling errors. It has been observed that with an increase in the size of the sample the unbiased errors decrease in magnitude.

Causes of error: Error may arise due to :

1) **Faulty process of selection:** Faulty selection of the sample may give rise to a bias in a number of ways such as:
   (a) Deliberate selection of a representative sample.
   (b) Conscious or unconscious bias in the selection of a random sample.
   (c) Substitution of a selected item in the sample by another

2) **Bias in analysis:** Faulty methods of analysis of data may also introduce bias in addition to bias which arises from faulty process of selection and faulty collection of information. Such bias can be avoided by adopting the proper methods of analysis.

3) **Bias due to faculty collection of data**: Bias may arise due to following reasons:

   a. Bias due to faculty collection of data: Bias may arise due to the following reasons: Unorganised collection procedure.
   b. Prejudice on the part of the person furnishing information.
   c. Prejudice of the person collecting information

Methods of reducing sampling errors:

The simplest way of reducing the sampling error is to increase the size of the sample. There is inverse relationship between the sampling error is very large when the size of the sample is small. The sampling error decreases with the increase of the sample size. When the sample increases in its size, it becomes more representative of the population. Sampling error reduced to zero when the size of the sample approaches the size of the population.

NON-SAMPLING ERRORS:

The errors which are arise in population are known as non-sampling errors. Sampling errors arise in the process of inferring about the population from the sample whereas non-sampling errors arise due to other causes.

The main causes of non-sampling errors are as follows:

(i) Vague definition of population.
(ii) Personal bias of the investigator.
(iii) Improper definition of the variables.
(iv) Application of wrong statistical methods.

(v) Errors in data processing. Errors committed during presentation and printing of tabulated results.

(vi) Errors due to non-response such as incomplete coverage in respect to units.

SAMPLING DISTRIBUTIONS

Samples of a given size may be drawn randomly from a population and the statistical e.g., mean, standard deviation, etc, may be computed for each such sample. The distribution of each such statistic is called the sampling distribution.

CHARACTERISTIC OF A SAMPLING DISTRIBUTION

Sampling Distribution is a frequency distribution renting the means taken from a great many samples, of the same size. The main characteristic of this is that is approaches normal distribution even when the population distribution is not normal provided the sample size is sufficiently large (greater than 30). The diagrams on the pages 27 and 28 will make the point clear.

THEORY OF ESTIMATION

From such sampling distribution it is possible to draw valid conclusions about the population parameters. A large part of the theory of sampling is devoted to finding from the sample estimates of the population parameters. Such population parameter may include the mean, dispersion, moments, and measures of skewness; and in multivariate population the various total and partial correlations.

In general, there are more ways than one of estimating a parameter from the data of a sample and the knowledge of the form of sampling distribution. Some of these ways will be better than others. The theory of estimation deals with these and related matters. It seeks to investigate the conditions which an estimate

should obey, what are the best estimates to employ in given circumstances, and how good are other estimates in comparison.
Obviously, estimates are bound to differ from the true value of the population parameters.

But the tolerable divergence between the estimates and the true value of the population parameter may be specified beforehand and the theory of sampling helps in determining the probability of such divergence. Put in another way, the degree of confidence that we can place in the estimate, can be established in terms of probability. As an example, if our aim was to determine the mean height of the soldiers in a regiment, we may take a sample of, say 100 soldiers and find the sample mean which may be something like 68". The theory of estimation helps us in making such statement as the population mean height of the regiment is likely to lie between 68"+1.5" with 95% of probability. The implication of this is that if repeated random samples of size 100 are taken from the regiment the means of 25% of such samples would be within the range 68"±1.5".

If 68"±11/2" is deemed to be a rather rough estimate only and it is desired to obtain the estimate in a closer interval, a large sample may have to be taken. The theory of estimation, then also helps in finding the size of the sample for the larger sample is required, the limit being the complete enumeration which would obviously yield the most precise measure of the parameter. The compromise, however, has to be struck between the increased costs associated with large samples and the satisfactory precision in estimation.

SIGNIFICANCE TESTING

Significance testing is concerned with ascertaining if the computed value of a statistic from a sample could have arisen from the given population. Consider for example, a company receiving large suppliers of an item from a vendor. It is known from past experience that the supplies contain 2% defectives. The

company is ready to accept 2% defectives but nothing more than this. It would be very expensive for the company to inspect each and every piece in the batches of supplies being received. Certain policies may be framed based on sampling distributions. Such rules may be used for routine, as “take a sample of 5 items; if there is no defective, accept the entire batch otherwise reject and return the entire batch”. These rules are ultimately based on Sampling distributions and ensure that, in the long run, the company receives supplies of 2% defectives with the desired probability. This is an easy and quick way of dispensing with the complete inspection of supplies which may range into thousands of items. The results of the sample data are being used for such decision –making. The underlying reasoning in that a sample of 5 items containing one or more defective could not possibly have come from supplies with an average of 2% defectives i.e., 1 or more defective in 5(2% defective in 5) (20% defectives) is significantly differ the acceptable average of 2% defectives.

Since the decision to reject or accept a batch is being based on the result of a sample, the decision is risky indeed. In the theory of estimation, explained earlier, the sampling as preferred to complete enumeration because it leads to quicker and cheaper estimation of population parameters, sacrificing some of the precision.

In significance testing, decisions are based on the sample results, because this is a more prompt and cheaper way although inherently it is risky. There are two categories of the risk involved. The particular batch was sufficiently good, but the sample had more than one defectives. This risk is being withstood by the supplier. In the long run a percentage of "good" batches is likely to be rejected. In the opposite case, a batch with more than 2% defectives may be accepted on the long run, the company should be accepting some percentage of batches which involves risk for both the supplier and the company; but with the advantage considerably reduced inspection expenses. The scheme should be so devised that

the two types of risk can be quantified beforehand and concurrence obtained from both the company and the supplier.

DISTRIBUTION OF SAMPLE MEANS

If all possible samples of size, n are taken from a population that is normally distributed, the sample means would also found to be normally distributed. The sampling distribution of the sample means would have the population mean as its mean, but the spread or dispersion would be lesser.

But even if the original distribution is non-normal and the sample size is fairly large (near 30), the distribution of the sample means tends to be normal. A fairly diverse forms of distribution lead to Normal distribution of the sample means provided n is approximately or larger than 30.

A precise statement of this law of large numbers is given by what is called the Central Limit Theorem.

CENTRAL LIMIT THEOREM

If a universe or population has a mean and a finite standard deviation, then the distribution the sample means approaches a Normal distribution with the mean μ and standard deviation, $\sigma/\sqrt{n}$ as the sample size increase.

Central limit theorem assures us that the sampling distribution of the mean approaches normal with the increase in the sample size. It permits us to use sample statistics for the purpose of making inferences about population parameters without having any knowledge about the shape of the frequency distribution of that population. The standard deviation of the sampling distribution $\sigma/\sqrt{n}$ is usually called the standard error.

STANDARD ERROR

Error in Statistics is different from what we mean by mistake in ordinary parlance. It is the difference of the actual value and the expected. Such error may arise on account of several factors and the methods of approximation, use of logarithms or the use of an average in place of an actual figure. The statistical error may also be due to improper definition of the variables, defective framing of questionnaire, application of wrong statistical methods, etc. These are all non-sampling statistical error while sampling errors are those which arise due to use of sampling as against complete enumeration. It is thus the measure of the divergence between the statistic and parameter values. This average once is likely to be the one which secures a compromise between the precision to be sacrificed and the effort in observing the samples of a given size.

The distribution of sample means being normally or approximately normally distributed about the population mean μ, the expected value of a sample mean is

$$E(\bar{x}) = \mu$$

The standard error of the mean is σ_x is the standard deviation of sample means, and is given by

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

n *b*eing the sample size.

Unbiased estimate of the standard error of the mean.

It may have to be estimated from the sample data. With large samples the sample standard deviation, s, may be substitute for σ, the population standard deviation for computing the standard error of the distribution of the sample, means.

A sample S.D is a biased estimated of population S.D. A good estimator shall be unbiased such that the expected value of the estimator is the value of the

parameter of all sizes. An unbiased estimate of population S.D. is either computer using $\sqrt{\sum \frac{(x-\bar{x})^2}{n-1}}$, or obtained by multiplying the sample S.D. by the fraction $\sqrt{n/(n-1)}$, n being the sample size. If n is sufficiently large (30 or more), the fraction $\sqrt{n/(n-1)}$ is very nearly equal to 1, and the different between the sample S.D. and the estimate of population s. d. is negligible (i.e. s=$\hat{\sigma}$). Therefore, the unbiased estimated of the population S.D. on the basis of the sample S.D. is given by$\hat{\sigma} = \frac{s\sqrt{n}}{n-1}$, where s stands for the sample s.d.

The unbiased estimate of standard error of the mean is given by

$\hat{\sigma} = \frac{\hat{\sigma}}{\sqrt{n}} = s \cdot \frac{\sqrt{n}}{\sqrt{n}-1} \cdot \frac{1}{\sqrt{n}} = \frac{s}{\sqrt{n}-1}$

The Finite Population Multiplier

The above mentioned expression for standard error of the mean is used when the population is either infinite or in which samples are drawn from a finite population with replacement.

In case, the population is finite, we use

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} x \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

Where N= size of the population, and n= size of the sample. The new term $\sqrt{N} - n\sqrt{N} - 1$ in the right- hand side is known as finite population multiplier.

The fraction $n\sqrt{N}$ is known as the sampling fraction. In case, the sampling fraction is small, the standard error of the mean for finite population is very close to the standard error of the mean for infinite populations. Because of this reason, the finite population multiplier is ignored when the sampling fraction is less than 0.05.

Estimation of Population parameters

An estimate is a specific observed value of a statistic. There are two types of estimates about a population parameter – 1) point estimate and 2) interval estimate.

POINT ESTIMATE

A point estimator is a single number that is used to estimate an unknown population parameter. It is useful if we have an idea of the error that might be involved.

Eg : Sample mean $\{\bar{x}\}$ is an estimator when population μ is unknown.

INTERVAL ESTIMATE

An interval estimate is a range of values used in making estimation of a population parameter. It has the advantage of showing the error in two ways – 1) by the extent of its range, and 2) by the probability of true population parameter lying within the range.

Interval estimates and confidence intervals

The probability that we associate with an interval estimate is called the confidence level. It indicates, how confidently we can say that the interval estimate will include the population parameter. The high the probability the more is the confidence. Although any confidence level may be considered, the most commonly include about 99% of the area under the curve.

From the table of area under the standard normal curve it is found that ± 1.64 standard errors include about 90% of the area under the curve; ± 1.96 standard errors include about 95% of the area under the curve; ± 2.33 standard errors include about 98% of the area under the curve; ± 2.58 standard errors include about 99% of the area under the curve.

CONFIDENCE INTERVAL ESTIMATE OF THE MEAN

On the assumption that the distribution of the sample mean is normal, a confidence interval estimate of the unknown population mean μ may be found by following the three steps given below:

Step 1. Select the confidence level and corresponding to that specific level of confidence note down the confidence coefficient z.

Step 2. If the population standard deviation σ is known, comute $\sigma_{\bar{x}}$ by using $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$..

If σ is unknown, but the sample standard deviation, s is known, compute an unbiased estimate of $\sigma_{\bar{x}}$ by using

$$\wedge_{\sigma_{\bar{x}}} = \frac{s}{\sqrt{n-1}}$$

n being the sample size.

Step 3. Construct a confidence interval as follows:

$$\frac{Known\ \sigma}{Population\ is\ normal - any\ n} \qquad \frac{Unknown\ \sigma}{Any\ population\ with\ large\ n}$$

Or

$$\frac{Any\ population\ with\ large\ n}{\bar{x} \pm z \wedge_{\sigma_{\bar{x}}}} \qquad \frac{Any\ population\ with\ small\ n}{\bar{x} \pm z \wedge_{\sigma_{\bar{x}}}}$$

From the table of areas under the Standard Normal Probability Distribution we form the following useful:

Confidence level	Confidence coefficient z
90%	1.64
95%	1.96
98%	2.33
99%	2.58
Without any reference to the confidence level	3.00

The confidence limits are the upper and lower limits of the confidence interval.

Example:A random sample of 50 items drawn from a particular population has a mean 30 with a standard deviation 28. Construct a 98% confidence interval estimate of the population mean.

Solution:

Step 1. Desired confidence level being 98%, the confidence coefficient, z, is 2.33.

Step 2. Since the population s.d. is unknown, we compute an unbiased estimate of the standard error of the mean, given by

$$\wedge_{\sigma_{\bar{x}}} = \frac{s}{\sqrt{n}-1} = \frac{28}{7} = 4$$

Step 3. Required confidence interval is given by $x+z\wedge_{\sigma_{\bar{x}}}$

i.e., $30 \pm 2.33x4$

i.e., 30 ± 9.32

Thus the upper confidence limit = 39.32

And the lower confidence limit = 20.68

Example. A pharmaceutical company wants to estimate the mean life of a particular drug under typical weather conditions. Following results were obtained from a simple random sample of 100 bottles of the drug:

Sample mean =	18 months
Population standard deviation =	5 months
Sample size =	100

Find an interval estimate with a confidence level of (a)90% (b)95% and (c)99%.

Solution: The sample size being large, we use the normal distribution as the sampling distribution. The standard error of the mean is given by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{100} = 0.5\ months$$

(a) 90% confidence interval is given

$\bar{x} \pm 1.64\sigma_x = 18 \pm (1.64x0.5) = 18 \pm 0.82$

Thus the upper confidence limit = 18.82 months;

And the lower confidence limit = 17.18 months.

(b) 95% confidence interval is given by

$\bar{x} \pm 1.96\sigma_x = 18 \pm (1.96x0.5) = 18 \pm 0.98$

Thus the upper confidence limit = 18.98 months;

And the lower confidence limit = 17.02 months.

(c) 99% confidence interval is given by

$\bar{x} \pm 2.58\sigma_x = 18 \pm (2.58x0.5) = 18 \pm 1.29$

Thus the upper confidence limit = 19.29 months;

And the lower confidence limit = 16.71 months.

Example: For the purpose of estimating the mean annual income x of 800 families of a particular community a simple random sample of size 56 was drawn and the following results were obtained:

$$\bar{x} = Rs.5{,}376$$

s = Rs. 840

Find a 90% confidence interval for the population mean.

Solution: The sample size bign large, we use the normal distribution as the sampling distribution.

Since we have a finite population of size 800, and since n/N = 56/800 = 0.07 is more than 0.05, using finite population multiplier, we have

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n-1}} \cdot \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

$$= \frac{840}{\sqrt{55}} \cdot \frac{\sqrt{800-56}}{\sqrt{800-1}} = \frac{840}{7.16} x0.9649$$

= Rs. 109.29

90% confidence interval is given by

$$\bar{x} + 1.64 \wedge_{\sigma_{\bar{x}}} = 5{,}376 \pm (1.64x109.29) = 5{,}376 \pm 179.23$$

Thus the upper confidence limit = Rs. 5,555.23

And the lower confidence limit = Rs. 5,196.77

Use of t distribution in interval estimation

"t" distribution was develop by W. S. Gossett, t distribution is used when the sample size is less, than 30 and the population standard deviation σ is unknown. It is to be noted that small sample size is only one of the conditions while using t distribution.

Assumptions for t-test

(i) Samples are drawn from normal population and are random.

(ii) For testing the quality of two population means population variances are regarded as equal.

(iii) In case of two samples some adjustment in degrees of freedom for t is made.

Properties of the t-distribution

(i) t- distribution is asymptotic to x-axis i.e., extends to infinity on either side.

(ii) t- distribution is symmetrical.

(iii) t- distribution has a greater spread than the Normal distribution. As n gets large, the t-distribution approaches the normal distribution. For $n \geq 30$ the two are fairly close.

(iv) The form of the t-distribution various with the degrees of freedom.

(v) The t-table may be found in any text. It gives, over a range of value of degrees of freedom (d.f.), the probabilities of exceeding by chance to value of t at different level of significance.

The quantity t is defined as

$$t = \frac{|Difference\ of\ means|}{S.E.of\ the\ mean}$$

With degrees of freedom n-1.

An interval estimate of the population mean is given by $\bar{x} \pm t \wedge_{\sigma_{\bar{x}}} t$ it is to be noted that the t-table does focus on the chance that the population parameter being estimated will not be within the desired confidence interval (i.e., it will be outside it). For making estimate at 95% confidence level the column under the head 05(100%-95%=5%) in the t-table is to be seen. We shall see the σ colum .05, .02, and .01 for confidence intervals of 95%, 98% and 99% respectively.

Degree of freedom

Degrees of freedom are the number of values we can choose freely. Let us consider two sample values, a and b, and let us suppose that they have a mean of 10.

i.e., $(a + b)/2 = 10 \Longrightarrow a + b = 20$

It suggests that a and b may be any two numbers such that a+b = 20. In case a is known; be is no longer a free number to take on any value. Thus, for a sample with two elements and known sample mean, we are free to specify only one of the elements. Again, if there are 5 elements in a sample, and if the mean is known we are free to specify 5-1 = 4 variables, are no longer free to specify the fifth variable, which is determined automatically.

Thus with 2 sample values we have 1 degree of freedom; with 5 sample values 4 degrees of freedom, and generalizing, with n sample values n-1 degrees of freedom. It is to be noted that there is a different t-distribution for each of the possible degrees of freedom.

Example: A sample of size 10 has a mean 37 with a standard deviation of 12. Construct a 99% confidence interval for the population mean.

Solution: Confidence level being 0.99, the column under the head 0.01 (Since 1-0.99 = 0.01) in the t-table for 10-1 = 9 d.f. is seen to be 3.250.

An unbiased estimate of the standard error of the mean is found to be

$\wedge_{\sigma_{\bar{x}}}$=s/ $\sqrt{n-1} = 12/3 = 4$

Required confidence interval for the population mean is given by

$$\bar{x} \pm t \wedge_{\sigma_{\bar{x}}} = 37 \pm 2.821x4 = 37 \pm 11.284$$

Thus the upper confidence limit = 48.284 and the lower confidence limit = 25.716.

Confidence Interval estimate of the proportion – large sample

Although we should consider the binomial distribution for constructing confidence interval estimate of a population proportion, it is a complex proposition, because, the computation of binomial probabilities is time consuming and tedious. But, binomial distribution can be approximated by an appropriate normal distribution as the sample size increase. If n be more than 30 and each of np be at least 5, we shall use the normal distribution is place of binomial. It is to be noted that the problem is concerned with the large samples.

We know that a population, p, is the ratio of the number of elements possessing a characteristic to the total number of elements in the population. If we multiply the proportion by 100, we obtain the percentage, and we may make use of percentage for the proportion and vice-versa.

A sample proportion, p, is the ratio of the number of elements possessing a characteristic to the total number of elements, n, in the sample. The mean of the sampling distribution of p equals to population proportion. The standard deviation of the sampling distribution of proportions i.e., the standard error of proportion is given by

$$\sigma_p = \frac{\sqrt{\overline{pq}}}{n}, where\ q = 1 - p$$

When p is unknown, and we have no other alternative but to make use of the sample proportion p, an unbiased estimate of the standard error of proportion

$\wedge_{\sigma_p = \sqrt{pq} \div \sqrt{(n-1)}}$

But if n is large, the expected different between the sample proportion and the population proportion approaches zero; i.e., for large samples, the sample proportion is an unbiased estimate of the population proportion.

The confidence interval estimate of the population, p, is given by either of the following:

Known population proportion, p	or	Unknown population proportion
$\bar{p} \pm z\sigma_p$		$\bar{p} \pm z \wedge_{\sigma_p}$

Note: The sample size being large, we shall use n in place of n-1 in the expression of, σ_p i.e. $\wedge_{\sigma_p} = \sqrt{\bar{p}\bar{q} \,/\, n}$

Example. A sample of 500 invoices is drawn randomly from a large population, and 36 percent were found to be incorrect. Construct a 90% confidence interval for the true proportion.

Solution: $\wedge_{\sigma_p} = \sqrt{\bar{p}\bar{q} \,/\, n} = \sqrt{36x(100-36)/500} = 2.146$

Required confidence interval based on percent = $\bar{p} \pm z \wedge_{\sigma_p}$

$= 36 \pm 1.64x2.146 = (36 \pm 3.52)\%$

Thus the upper confidence limit = 32.48%

And the lower confidence limit = 39.52%

If we make use of the proportions, the computation will be as follows:

$$\wedge_{\sigma_p} = \sqrt{0.36x(1-0.36) \,/\, 500} = 0.021$$

Required confidence interval = 0.36±1.64x0.021 = 0.36±0.034

Thus the confidence limits are 0.326 and 0.394

Example: In a locality containing 18,000 families a sample of 840 families was selected at random. Of these 840 families, 206 families were found to have a monthly income of Rs.50 or less. It is desired to know how many out of 18,000 families have a montly income of Rs.50 or less.

Solution: p, the proportion of families having a monthly income of Rs.50/- or less = 206/840 = 0.245 and q, the proportion of families not having a monthly income of Rs.50/- or less = 1-(206/840)= 634/840.

$$\wedge_{\sigma_p} = \sqrt{\frac{pq}{n}} =$$

Since there is no reference to the confidence level. 3 sigma limits for the population proportion are $0.245 \pm 3x0.245 \pm 0.045$.

Thus the upper confidence limit = 0.29

and the lower confidence limit = 0.20

Out of 18,000 families 18,000 x 0.245 = 4410 families are likely to have income of less than Rs. 50/-

The limits would be 18,000 x 20/100 = 3,600 and 29/100 = 5,220.

SAMPLING IN ACCOUNTING AND AUDITING : The application in accounting includes audit tests, estimation of value of inventories, estimation of intercompany transaction, control of clerical errors, costing operations, etc.

The auditor is often confronted with a large number of documents and entries which are too many for him to perform a 100 percent examination in the available time and at a reasonable cost. He may then have selective test. In order, therefore, to satisfy himself of the adequacy or reliability of check as also for proper generalization of results, it is necessary to apply scientific sampling techniques. He can specify and objectively state the risk he is willing to take for that the sampling estimate will not be as accurate as he desired. Having stated

these specification i.e., accuracy and risk, a sample size can then be computed which will achieve the auditor's objectives.

Determination of Sample Size

It may be noted that if the size of the sample is too small it may not help in the analysis: on the other hand if it is too large, there may be waste of resources. To strike a balance between the two we consider the problem of determination of right sample size with a specified level of precision.

SAMPLE SIZE FOR ESTIMATING A MEAN

In determining the sample size for estimating a population mean, following three factors must be known:

1) the desired confidence level
2) the permissible sampling error, E= $\bar{x} - \mu$, and
3) the standard deviation, σ.

Having the knowledge of above three factors,

N = ($\sigma\ z\ /\ E$)^2 = (1096x20/5)^2 = 61.46

Therefore, the size of the sample is 62.

SAMPLE SIZE FOR ESTIMATE A PROPORTION

The procedures for determining the sample size for estimating a population are similar to those of estimating a population mean In this case, also, we must known the following three factors.

1) The desired confidence level,
2) The permissible sampling error, E.
   (i.e. the difference between the estimate from the sample p and the parameter to be estimated p). and
3) The estimated true proportion of success.

The sample size n is given by n = ($z^2pq\ /\ E^2$), $where\ q = 1 - p$.

Unit-IV

Testing Hypothesis - one sample tests – Hypothesis testing of mean when the population standard deviation is known, powers of hypothesis test, hypothesis testing of proportions, Hypothesis testing of means when standard deviation is not known.

Testing Hypothesis – Two sample tests - Tests for difference between means - large sample, small sample, with dependent samples, testing for difference between proportions, Large sample.

TESTING OF HYPOTHESES

A Hypothesis is a statement about the population parameter. Hypothesis testing is a procedure that helps us to ascertain the likelihood of hypothesized parameter being correct by making use of the sample statistics. The two hypotheses in a statistical test are normally referred to as,

(i) Null Hypothesis

(ii) Alternative Hypothesis.

(i) Null Hypothesis: Null Hypothesis which is tested to be actually tested for acceptance or rejection is termed as Null Hypothesis. According to R A Fisher, "Null Hypothesis is the hypothesis which is tested for rejection under the assumption that it is true".

The null hypothesis is a very useful tool to test the significance of difference. In the process of statistical test, the hypothesis is rejected or accepted based on the sample drawn from the population. This hypothesis reveals that the mean of the sample and the mean of the population under study do not show any difference.

A statistical hypothesis is a null hypothesis if it is accepted. We should take into consideration the following while setting up the null hypothesis:

a) To test the significance of the difference between the values of the sample and the population, or between two sample values; we set up the null hypothesis that the difference is not significant. This is because the difference is due to sample fluctuations.

$H_0 : \mu = \bar{x}$

Where μ = population mean

$\bar{x}$ = sample mean

b) To test any statement about the population, we set up the null hypothesis, that it is true.

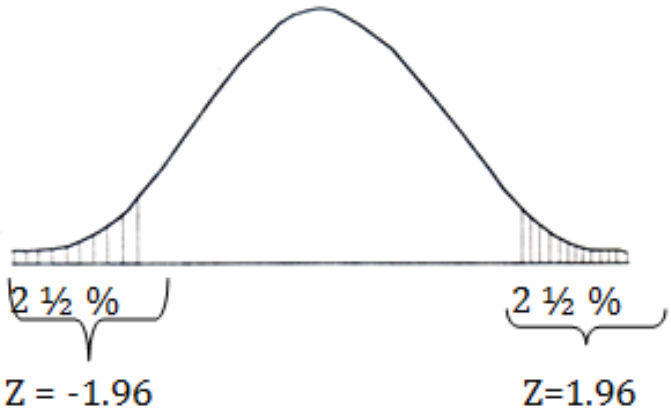
(ii) Alternative Hypothesis: "Any hypothesis which is complimentary to the null hypothesis is called an alternative hypothesis.". Rejection of H_0 leads to the acceptance of alternative hypothesis which is denoted by H_1.

For example, if we want to test the null hypothesis for difference between population mean and sample mean then we these hypothesis can be written as follows:

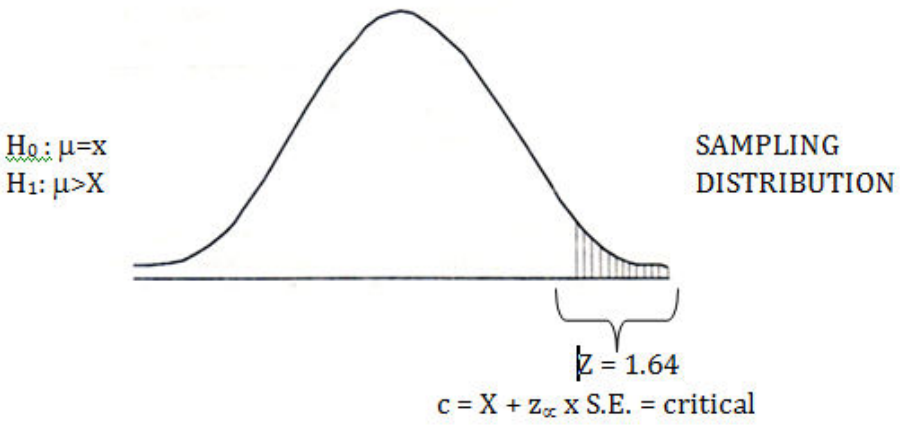
$H_0 : \mu = \bar{x}$ (Null Hypothesis)

$H_1 : \mu \neq \bar{x}$ (two-tailed alternative hypothesis)

$H_1 : \mu > \bar{x}$ or $H_1 : \mu < \bar{x}$ (right tailed and left tailed tests)

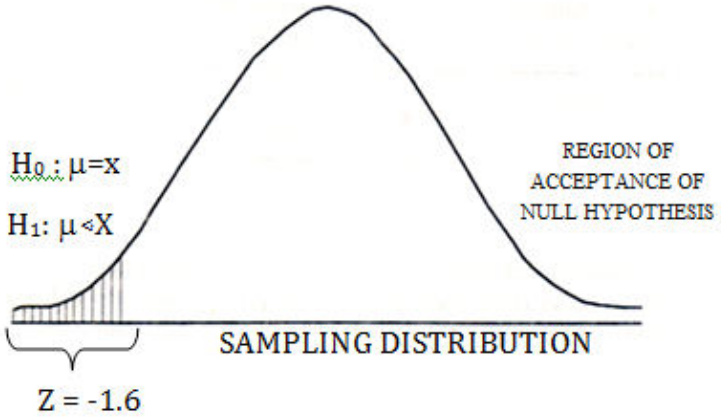


Two-tailed diagram



Value for upper time

c = X - zσ x S.E. = critical value for left tail



The validity of H_0 and H_1 is then ascertained at a certain level of significance. The significance level stands for the confidence with which the experimenter rejects or retains the Null Hypothesis. Significance level is customarily expressed as a percentage e.g., 5% significance implies the probability of rejecting the Null hypothesis is I in 20, when it is true.

Having set up the Null and the Alternative Hypotheses, and the significance level, the next step is to construct a test criterion. This involves selecting the right probability distribution for the particular test.

Following table shows the conditions for using z-test and t-test in testing hypothesis about means:

Sample size, n	Population s.d., σ is known	Population s.d.,σ is unknown
n > 30	z-test	z-test
n < 30	z-test	t-test

To perform computation on the sample statistic and ascertain if the computed value falls in the region of rejection or of acceptance (i.e., non rejection), based on the chosen level of significance. The value of the sample statistic that defines the regions of acceptance and rejection is referred to as the *critical value.* It may be noted that the critical region appears in one or both tails of the sampling distribution of the test statistic. The area (giving the probability) in the tails equals the level of significance α . For one-tailed test α appears in one tail, and for two-tailed test a/2 appears in each tail of the distribution.

ERRORS IN HYPOTHESIS TESTING

In any hypothesis testing procedure there is always a risk of arriving at an incorrect conclusion.

Associated with any test there are two types of errors -

1) The Type I error
2) The Type II error, and these are related to the null hypothesis.

Type I error: Reject H_0 when it is true. The probability of making a Type I error is given by α, the level of significance.

Type II error: Accept H_o, when it is false. The probability of making a Type II**,** it is denoted β.

The probability of making one type of error can be reduced only by allowing an increase in the probability of other type of error. The trade-off between these two types of error is made by assigning appropriate significance level after examining the costs or penalties attached to both types of error.

PROCEDURE FOR TESTING OF HYPOTHESIS:

1) Null hypothesis : Set up the null hypothesis = H_0
2) Alternative hypothesis: Set up the Alternative hypothesis = H_1
3) This will enable us to decide whether we have to use a single tailed (right or left) test or two tailed test.
4) Level of significance: Choose the appropriate level of significance α
5) Test statistic: Calculate the value of z (the test statistic) under the null hypothesis.
6) $|z| = \frac{\bar{x} - E(\bar{x})}{S.E(\bar{x})} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
7) Conclusion: Compare the calculated z value with the tabulated value, at t he given level of significance α.

   If $|z| > z_\alpha$ (the table value)

We say that it is not significant, that means the difference x -μ is just due to fluctuations of sampling and the sample data do not provide us sufficient evidence against the null hypothesis which may be accepted.

If $|z| > z_\alpha$ i.e. calculated |z| value is greater than the tabulated value, we say that it is significant and the null hypothesis is rejected at the level of significance α.

Test of significance for large samples:

We know that for large values of n, the number of trials, binomial. Poisson are very closely approximated by normal distribution.

If $X \sim N(\mu, \sigma^2)$ and $z = \frac{\bar{X} - \mu}{\sigma} = \frac{X - E(X)}{Var(X)} \sim N(0,1)$

From the normal probability tables

$P(-3 \leq z \leq 3) = 0.9973$ i.e., $P(|z| \leq 3) = 0.9973$

$\Rightarrow P(|z| \leq 3) = 1 - P((|z| \leq 3) = 0.0027$

$\therefore |z| \leq \pm 3$

From the tables

$P(-1.96 \leq z \leq 1.96) = .95$ i.e. $P(|z| \leq 1.96) = 0.95$

Similarly $P((|z| \leq 2.58) = 0.99$

$P(|z| > 2.58) = 0.01$

$\therefore$ The significant values of z at 5% and 1 % level of significance from a two tailed test are 1.96 and 2.58 respectively.

If $|z| > 3$, H_0 is rejected

If $|z| < 3$, we test its significance at certain level of significance.

If $|z| > 1.96$, H_0 is rejected at 5% level of significance.

If $|z| > 2.58$, H_0 is rejected at 1% level of significance.

$P(z > 1.645) = 0.5 - P(0 \leq z \leq 1.645)$

$= 5 - 0.45 = 0.05$

$P(z > 2.33) = 0.5 - P(0 \leq z \leq 2.33)$

$= 0.5 - 0.459 = 0.01$

For single tailed test (for right tailed test) if $z > 1.645$. H_0 is rejected at 5% level.

If $|z| > 2.33$, H_0 is rejected at 1% level for right tailed test.

Test of significance for single mean:

We know that if x_i (i=1,2,3…n) is a random sample of size n taken from a normal population with mean μ and variance σ^2, then the sample mean is distributed normally with mean μ and variance $\frac{\sigma^2}{n}$ i.e., $x \sim N\left(\mu \frac{\sigma^2}{n}\right)$. However, this result holds i.e., $x\ N\left(\mu, \frac{\sigma^2}{n}\right)$ even from random sample taken from a non – normal population provided the sample size n is large (central limit theorem)

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

For small samples (σ unknown) sample variance can be calculated by

$$S^2 = \sum_{i=1}^{n} \frac{\sum (x_i - x)^2}{n-1}$$

Null hypothesis = $\mu = \mu_0$ and test statistic (for single mean) = $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$

Which is a random variable having test t – distribution with (n -1) degree of freedom.

1) **Example:** According to the norms established for a mechanical aptitude test, persons who are 18 years old have an average height of 73.2 with a standard deviation of 8.6. If 45 randomly selected persons of that age averaged 76.7. Test the null hypothesis μ=73.2 against the alternative hypothesis $\mu > 73.2$ at the 0.01 level of significance.

Solution: Null hypothesis : $\mu = 73.2$

Alternative hypothesis $\mu > 73.2$ (Right tailed test)

Level of significance = 99% or probability is .01

Test statistic: $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

$\mu = 73.2$

x = mean of the sample = 76.7

σ = standard deviation of population = 8.6

N = sample size = 45

$Z = \frac{76.7-73.2}{\frac{8.6}{\sqrt{45}}} = 2.73$

Table value $z_\alpha = 2.33$

Calculated value is greater than table value

$\therefore$ H_0 is rejected.

2) Tests performed with a random sample of 40 does; engines produced by a large manufacturer show that they have a mean thermal efficiency of 31.45% with a standard deviation of 1.6%. At the .01 level of significance. Test the null hypothesis μ =32.3% against the alternative hypothesis $\mu \neq$ 32.3%

Solution: Null hypothesis $\mu = 32.3\%$

Alternative hypothesis $\mu \neq 32.3\%$

Level of significance = .01

Table value of z = 2.58 (two tailed test)

Test statistic z = $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

x = mean of sample = 31.45%

μ = 32.3

σ = standard deviation = 1.6%

n = sample size = 40

z = $\frac{31.45-32.3}{\frac{1.6}{\sqrt{40}}} = \frac{-.85}{\frac{1.6}{\sqrt{40}}}$ = -3.33

|z| > 3

If |z| > 3, H_0 should be rejected.

$\therefore$ H_0 is rejected

$\therefore$ $\mu \neq 32.3$

Test of significance for difference of means :

To test the significance for difference of means. Suppose we select two samples which are independent to test the significance.

Let $\overline{x_1}$ be the mean of a simple of size n_1 drawn from a population of mean μ_1 and variance σ_1^2 and $\overline{x_2}$ be the mean of an independent random sample of size n_2 drawn from another population with mean μ_2 and variance σ_2^2 , since n_1 and n_2 are large.

$$\overline{x_1} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \overline{x_2} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

$(\overline{x_1} - \overline{x_2})$ is the difference between two independent normal variates and hence also a normal variate, z for ($\overline{x_1}$ - $\overline{x_2}$).

We want to test the null hypothesis $\mu_1 - \mu_2 = \delta$ where δ is a specified constant. μ_2 and the variances σ_1^2 and σ_2^2, then the distribution of their sum (or difference has the mean $\mu_1 + \mu_2$ or $(\mu_2 - \mu_1)$ and the variances $\sigma_1^2 + \sigma_2^2$

In testing the significance for the difference of two means, we shall consider the alternative hypothesis $\mu_1 - \mu_2 < \delta$, $\mu_1 - \mu_2 > \delta$ or $\mu_1 - \mu_2 \neq \delta$

To find the variances of the difference between the means of two samples

$$\sigma x_1^{-2} = \frac{\sigma_1^2}{n_1}, \sigma x_2^2 = \frac{\sigma_2^2}{n_2}$$

$$\sigma^2(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - E(x_1 - x_2)}{S.E(\bar{x}_1 - \bar{x}_2)}$$

Case i) under the null hypothesis $H_0 = \mu_1 + \mu_2$ there is no significant difference between the sample means $E(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2 = 0$

$$\therefore z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Case ii) if the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta$ or $\mu_1 - \mu_{2 < \delta}$ or $\mu_1 - \mu_2 > \delta$ i.e. $\mu_1 - \mu_2 \neq 0$

$$z = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Critical region for testing for testing $\mu_1 - \mu_2 = \delta$

Alternative hypothesis	Reject null hypothesis if
$\mu_1 - \mu_2 < \delta$	$z < -z_\alpha$
$\mu_1 - \mu_2 > \delta$	$z > -z_\alpha$
$\mu_1 - \mu_2 \neq \delta$	$Z < -z_{\alpha/2}$ $Z > z_{\alpha/2}$

When $H_0 : \mu_1 - \mu_2$ i.e., $\delta = 0$

Alternative hypothesis	Reject null hypothesis if
$\mu_1 < \mu_2$	$z \leq -z_\alpha$
$\mu_1 > \mu_2$	$z > -z_\alpha$
$\mu_1 \neq \mu_2$	$Z < -z_{\alpha/2}$ $Z > z_{\alpha/2}$

Note: If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $H_0 : \mu_1 = \mu_2$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Worked out problems

Example 1 : An investigation of two kinds of photocopying equipment showed that 71 failures of the first kind of equipment took on the average 83.2 minutes to repair with a standard deviation of 19.3 minutes, while 75 failures of the second kind equipment took on the average 90.8 minutes to repair with a standard deviation of 21.4 minutes. Test the null hypothesis $\mu_1 - \mu_2 \neq 0$ at the level of significationce $\alpha = .05$

Solution: Null hypothesis $\mu_1 - \mu_2 = 0$

Alternative hypothesis $\mu_1 - \mu_2 \neq 0$

Level of significance $\alpha = .05$

Test statistic $z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

$\bar{x}$ = mean of first sample = 83.2 minutes

$\bar{y}$ = mean of second sample = 90.8 minutes

σ_1^2 = variance of 1st sample = $(19.3)^2$sq. minutes

σ_2^2 = variance of second sample = $(21.4)^2$ sq. minutes

n_1 = size of first sample = 71

n_2 = size of second sample = 75

$z = \frac{83.2 - 90.8}{\sqrt{\frac{372.5}{71} + \frac{457.96}{75}}} = \frac{-7.6}{\sqrt{11.36}} = 2.2$

Conclusion:

$z_{\alpha/2}$ for .05 level is 1.96

$z > z_\alpha$

$\therefore$ Null hypothesis is rejected.

$\therefore \mu_1 \neq \mu_2$

$\therefore$ It does not take equal amount of time to repair either kind of equipment.

Example 2 Suppose that we want to investigate whether on the average man earn more than 20 per week more than women in a certain industry. If sample data show that 60 men earn on the average $\bar{x}_1$ = 292.50 per week W.a S.D. of 15.6 while 60 women carn on average y = 266.10 per week wa S.D of 18.20. What can we conclude at the .0 level of significance.

Solution : Null hypothesis : $\mu_1 - \delta = \mu_2$

Alternative hypothesis $\mu_1 - \delta > \mu_2$

Level of significance = .01

Test statistic : $z = \frac{\bar{x} - \bar{y} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

X = mean earn of men = 292.5

Y = mean earn of women = 266.10

σ_1 = standard deviation of earn of men = 15.6

σ_2 = standard deviation of earn of women = 18.2

N_1 = 60 = sample size of I sample i.e. No. of men

N_2 = number of women = 60

$Z = \frac{292.5 - 266.1 - 20}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}}$ (since $n_1 = n_2 = n$)

$= \frac{6.4}{\sqrt{\frac{1}{60}(243.4 + 331.2)}} = \frac{6.4}{3.1} = 2.07$

Conclusion : $z_\alpha = 2.33$ for .01 level of significance this is a right tailed test.

$|z| < z_\alpha$

$\therefore$ Null hypothesis should be accepted.

Test of significance for single proportion

If X is the number that an event occurs among n trials. The proportion of the time that the even actually occurs i.e., $\frac{X}{n}$ is the sample proportion. If n trials satisfy the assumptions underlying the binomial distribution, we know that the mean and standard deviation of the number of success are given by np and $\sqrt{np(1-p)}$ where p is the probability of success.

i.e., E(X) = np, V(X) = np (1-p). [q = 1-p is the probability of failure]

$\therefore \sigma = \sqrt{np(1-p)}$

If we divide these by E(X) and σ by n

We get $E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}.np = p$

$\left(\frac{\sigma}{n}\right) = \frac{\sqrt{np(1-p)}}{n}$

$\therefore$ Sample proportion is an biased estimator of the binomial parameter p. If we write P for probability of success and Q for probability of failure i.e., Q=1-P

P = sample proportion.

$\therefore E(p) = E\left(\frac{X}{n}\right) = P$

$V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{nPQ}{n^2} = \frac{PQ}{n}$

$\therefore$ standard error of $p = \sqrt{\frac{PQ}{n}}$

ii) If the sample is taken from a finite population of size N then standard error of proportions

$$S.E.(p) = \sqrt{\frac{N-n}{N-1}.\frac{PQ}{n}}$$

To find the confidence interval for p

Having approximately the degree of confidence $(1-\alpha)100\%$.

Suppose x_0 is the largest integer for which the binomial probabilities

b (k, n, p) = satisfy $\sum_{k=0}^{x_0} b(k,n,p) \leq \frac{\alpha}{2}$

Now x_0 and x_1 depend upon the value of P we can write x_0 (P),x_1(P)

We can assert with a probability of approximately 1 - α and at least 1-α

$\therefore$ for large n

$Z = \frac{X-E(X)}{V(X)} = \frac{X-nP}{\sqrt{nPQ}}$ confidence limits for p interms of the observed value X substituting $\frac{X}{n}$ for p

$$\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

is the confidence interval for p. (proportions)

if we write $\frac{x}{n}$ =P

The confidence interval for large sample for P

$$p - z_{\alpha/2}\sqrt{\frac{PQ}{n}} < p < p + z_{\alpha/2}\sqrt{\frac{PQ}{n}} \left(P = \frac{x}{n}, Q = 1 - P\right)$$

Maximum error:

The magnitude of the error, we make, when we use $\frac{x}{n}$ as an estimator of p is given by $\left|\frac{x}{n} - p\right|$.

Using the normal approximation, we can assert with probability 1 - α

$\therefore$ The inequality $\left|\frac{x}{n} - p\right| \leq z_{\alpha/2}\sqrt{\frac{PQ}{n}}$ will be satisfied.

$\therefore$ The error will be at the most $z_{\alpha/2}\sqrt{\frac{PQ}{n}}$

$\therefore$ Maximum error of estimate for the proportion p

$$E = z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

Sample Size:

Maximum error of estimate for the proportion p

$$E = z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

$$\therefore \sqrt{n} = z_{\alpha/2}\sqrt{\frac{PQ}{n}}$$

$$\therefore n = \left[\frac{\frac{z\alpha}{2}}{E}\right]^2 (PQ)$$

$\therefore$ Sample size n = (PQ) $\left[\frac{z_{\alpha/2}}{E}\right]^2$ or $\left[\frac{z_{\alpha/2}}{E}\right]^2$ P(1-P)

Note: If P is not given, we cannot use this formula. If p is not given; we can make

Use of the fact that P(1-P) is at the most $\frac{1}{4}$ (i.e., P=$\frac{1}{2}$)

$\therefore$ Sample size n (when p is not given) = $\frac{1}{4}\left[\frac{z_{\alpha/2}}{E}\right]^2$

One sided confidence interval for P:

If P < C where C is a constant depending on the degree of confidence and the size of the sample.

We know that, the binomial distribution is best approximated with a poisson distribution with $\mu = n P$ when P is small and n is large, we have,

P < $\frac{1}{2n}x_\alpha^2$ where $x_\alpha^2 = \frac{(n-1)S^2}{\sigma^2}$ with P 2(x +1) degrees of freedom).

6.5 Hypothesis concerning one proportion

We shall test the null hypothesis $P = P_0$ against one of the alternatives $P<P_0$ or $P > P_0$ or $P \neq P_0$ statistic for large sample test concerning proportion P which is a random variable having approximately the standard normal distribution. The initial regions are same as μ (usually we test at the level of .01 and .05 significance)

Worked out problems

1. In a random sample of 200 claims filed against an insurance company writing collision insurance on cars 84 exceeds 1200, construct a 95% confidence interval for the true proportion of claims filed against this insurance company that exceed 1200.

Solution: P = $\frac{x}{n} = \frac{84}{200}$ = .42, n = size of the sample = 200

Q = 1-P = $\frac{116}{200}$ = .58

Confidence interval for p.

$P - z_{\alpha/2}\sqrt{\frac{PQ}{n}} < p < + z_{\alpha/2}\sqrt{\frac{PQ}{n}}$

$z_{\alpha/2} = 1.96$

$0.42 - 1.96\sqrt{\frac{.42x\ .58}{200}} < p < 42 + 1.96\sqrt{\frac{.42x\ .58}{200}}$

$= 0.42 - 0.068 < p < 0.42 + 0.068$

$0\ .352 < p < 0\ .488$

2. What can we say with 99% confidence about the maximum error, if we use the sample proportion as an estimate of the true proportion of claims field against this insurance company that exceeds 1200 in the above problem.

Solution: Maximum error $E = z_{\alpha/2}\sqrt{\frac{PQ}{n}}$

$z_{\alpha/2}$ for (0.99) 100% confidence = 2.58

$+ = 0\ .42$, $Q = 58$, $n = 200$

$\therefore$ Maximum error $E = 2.58\sqrt{\frac{.42\ x\ 5.8}{200}} = 0\ .09$

3. In a random sample of 400 industrial accidents it was found that 23 I were due to least partially to unsafe working conditions. Construct a 99% confidence interval for the corresponding true proportion using large sample confidence formula.

Solution: Confidence interval $P - z_{\alpha/2}\sqrt{\frac{PQ}{n}} < p < p + z_{\alpha/2}\sqrt{\frac{PQ}{n}}$

P = probability of success = $\frac{231}{400} = .578$

$Q = 1 - P = 0\ .4122$

$z_{\alpha/2} = 2.58$, $n = 400$

$z_{\alpha/2}\sqrt{\frac{PQ}{n}} = 2.58\sqrt{\frac{.422\ x\ .578}{400}} = .064$

$\therefore$ Confidence interval $(0.578 - 0.064, 0\ .578 + 0.064) = (0.514, 0.642)$

3. In the above problem what can we say with 95% confidence about the maximum error if we use the sample proportion to estimate the corresponding true proportion.

Solution: P = 0.578, Q= 0.422

N = 400

$z_{\alpha/2} = 1.96$

Maximum error = $z_{\alpha/2}\sqrt{\frac{PQ}{n}} = 1.96\sqrt{\frac{.578 \times .422}{400}} = .048$

$\therefore$ Maximum error E = 0.048

Test of significance for difference of proportions.

Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute, say A, among their members. Let X_1, X_2 be the number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 taken from two populations respectively. Then sample proportions are given by:

$p_1 = \frac{X_1}{n_1}$, $p_2 = \frac{X_2}{n_2}$

If p_1 and p_2 are population proportions then

$E(p_1) = P_1$, $E(p_2) = P_2$

$V(p_1) = \frac{P_1Q_1}{n_1}$ $V(p_2) = \frac{P_2Q_2}{n_2}$

since for larger samples, p_1 and p_2 are asymptotically normally distributed (p_1-p_2) is also normally distributed.

The standard variable corresponding to the difference $(p_1 - p_2)$ is given by

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0, 1)$$

Under the null hypothesis

$H_0 : P_1 = P_2$ i.e. there is no significant difference between the sample proportions

$$\therefore E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2 = 0$$

$$V(p_1 - p_2) = V(p_1) + V(p_2)$$

$$\therefore V(p_1 - p_2) = \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2} = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Since under $H_0 = P_1 = P_2 = P.Q_1 = Q_2 = Q$

$\therefore$ Under $H_0 P_1 = P_2$

$$Z = \frac{P1 - P2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

And $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$

$$E(P) = \frac{1}{n_1 + n_2} E[n_1 p_1 + n_2 p_2]$$

$$= \frac{1}{n_1 + n_2} [n_1 E(p_1) + n_2 E(p_2)]$$

$$= \frac{1}{n_1 + n_2} [n_1 p_1 + n_2 p_2] = P \quad \therefore P_1 = P_2 = P \text{ under } H_0$$

$\therefore$ The estimate is unbiased

Large sample confidence interval for the difference of two proportions

$$= (p_1 - p_2) \pm z_{\alpha/2} \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Standard error of $(p_1 - p_2)$ - $\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Maximum error = $z_{\alpha/2} \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Note : Suppose the population proportions P_1 and P_2 are given to be different i.e., $P_1 \neq P_2$ and we want to test whether the difference $(P_1 - P_2)$ is significant then the test statistic becomes

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{S.E.(p_1 - p_2)} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

Here the sample proportions are not given. If we set up the null hypothesis $H_0 : p_1 = p_2$, the difference in population proportions is likely to be hidden in sampling. Then the test statistic becomes

$$| z | = \frac{|p_1 - p_2|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

Confidence interval = $(p_1 - p_2) \pm z_{\alpha/2} = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

S.E. $(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

And maximum error = $z_{\alpha/2} \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

1. The owner of a machine shop must decide which of two snack vending machines to install in his shop. If each machine is tested 250 times, the first machine fails to work 13 times and the second machine fails to work 7 times test at the 0.05 level of significance whether the difference between the corresponding sample proportions is significant.

Solution: Null hypothesis : $p_1 = p_2$

Alternative hypothesis $p_1 \neq p_2$

Level of significance $\alpha = .05$

Test statistic = $z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$p_1 = \frac{X_1}{n_1} = \frac{13}{250}$, $n_1 = 250$, $n_2 = 250$ $p_2 = \frac{X_2}{n_2} = \frac{7}{250}$

$P = \frac{X_1 + X_2}{n_1 + n_2} = \frac{13+7}{250+250} = \frac{20}{500} = .04$

$Q = 1 - P = 0.96$

$$Z = \frac{\frac{13}{250} - \frac{7}{250}}{\sqrt{.96 \text{ x } .04\left(\frac{1}{250}+\frac{1}{250}\right)}} = 1.37$$

Table value $z_{\alpha/2} = 1.96$

Conclusion = $z < z_{\alpha}$

$\therefore$ H_0 is accepted

$\therefore$ There is no significant difference between the sample proportions.

2. Photolithography pays a central role in manufacturing integrated circuits made on thin discs of silicon prior to a quality improvement program, too many rework operations were required. In a sample of 200 units, 26 required reworking of the photo lithographic step. Following training in the use of pareto charts and other approaches to identify significant problems, improvements were made. A new sample of size 200 had only 12 that needed rework. Find a large sample 99% confidence interval for the difference of the true proportions.

Solution: Confidence interval for the difference of two proportions is

$$(p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$p_1 = \frac{x_1}{n_1} = \frac{26}{200} = .13$, $p_2 = \frac{x_2}{n_2} = \frac{12}{200} = .06$, $z_{\alpha/2} = 2.58$

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{.13x\ .87}{200} + \frac{.06\ x.94}{200}} = 0.029$$

$$P_1 - p_2 \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = 0.13\text{-}0.06 \pm 2.58 \text{ x } 0.029$$

2.58 x0 .029 = 0.075

(0.07 -0.075, 0.07 +0.075)

$\therefore$ confidence Interval = (-0.005, 0.145)

Test of significance for the difference of Means (σ unknown) :

Under the null hypothesis (H_0) the sample have been drawn from the normal populations with means μ_x and μ_y and under the assumptions that the population variances are equal i.e., $\sigma x^2 = \sigma y^2 = \sigma^2$

The test statistic $t = \frac{\bar{x}-\bar{y}-\left(\mu_x-\mu_y\right)}{S\sqrt{\frac{1}{N}+\frac{1}{N_2}}}$

Where $\bar{x}$ = mean of 1st sample

$\bar{y}$ = mean of second sample

$S^2 = \frac{1}{(n_1+n_2-2)}\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2+\sum_{j=1}^{n_2}(y_i-\bar{y})^2\right]$

Since $\mu_x = \mu_y$ (under H_0)

$T = \frac{\bar{x}-\bar{y}}{S\sqrt{\frac{1}{n}+\frac{1}{n_2}}}$

Degrees of freedom = $n + n_2 - 2$

Confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x})_2 \pm t_{\alpha/2}\sqrt{\frac{\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2+\sum_{i=1}^{n_2}(y_i-\bar{y})^2\right]}{(n_1+n_2-2)}} x \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

1. A trucking firm suspects the claim that the average life time of certain tires is at least 28,000 miles. To check the claim the firm puts 40 of these tires on its trucks and gets a mean life time of 27463 miles with a standard deviation of 1348 miles. What can it conclude it the probability of a type 1 error is to be at most 0.01?

Solution : Null hypothesis μ= 28000 miles

Alternative hypothesis μ<28,000 miles. (left tailed test)

Level of significance α = .01

Test statistic $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

X = mean of sample = 27463

μ = mean life time of population = 28000

σ = standard deviation. Here S.D. of sample is given = 1348

N = 40

Z = $\frac{27463-28000}{\frac{1348}{\sqrt{40}}}$ = -2.52

Conclusion: z_α = -2.33

$Z < -2.33$

$\therefore$ Null hypothesis is rejected at .01

$\therefore$ $\mu < 28{,}000$ i.e. Alternative hypothesis should be accepted.

TEST OF SIGNIFICANCE (SMALL SAMPLES)

The Probability density function of t – Distribution

$$f(t) = \frac{r\left(\frac{v+1}{2}\right)}{\sqrt{v}\sqrt{\pi}\, r\left(\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-\left(\frac{v+1}{2}\right)} \text{ for } -\infty < t < \infty$$

The graph is similar to normal curve

Applications of t – distribution:

1) To test the significance of the mean of a small radom sample from a normal population.
2) To test the significance of the difference between the means of two samples taken from a normal population
3) To test the significance of an observed coefficient of correlation including partial and rank correlation.
4) To test the significance of an observed regression coefficient.

Eg 1. A random sample of six steel beams has a mean compressive strength of 58392 with a standard deviation of 648. Use this information at the level of significance $\alpha = .05$ to test whether the true average compressive strength of the steel from which this sample came is 58,000.

Solution: Null hypothesis $\mu = 58{,}000$

Alternative hypothesis = $\mu \neq 58{,}000$

Level of significance $\alpha = .05$

Test statistic $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

$\bar{x}$ = sample mean = 58392

S = standard deviation of the sample = 648

μ = 58,000 – Mean of the population to be tested.

n>30. This is small sample with 5 degrees of freedom.

Table value for α = .05 and v = 5 = 2.57

$t = \frac{58392-58000}{\frac{648}{\sqrt{6}}} = 1.49$

$t < t_{\alpha/2}$ $\therefore$ Null hypothesis accepted

$\therefore$ Mean of the population μ = 58,000

2. Given a random sample of 5 pints from different product lots. We want to test whether the fat content of a certain kind of ice cream exceeds 14% what can we conclude at .01 level of significance about the null hypothesis μ -14% if the sample has the mean $\bar{x}$ = 14.9% and the standard deviation s=42%.

Solution : Null hypothesis μ = 14%

Alternative hypothesis μ > 14% Right tailed test.

Test statistic $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

X= mean of the sample = 14.9%

μ = mean of the population μ = 14%

S = standard deviation of the sample = 42

t_α = +3.75

$t = \frac{14.9-14}{\frac{.42}{\sqrt{5}}} = 4.8$

$t > t_\alpha$

$\therefore$ H_0 is rejected. AceptH_1

$\therefore$ Mean of the population μ > 14%

3. A random sample from a company's very extensive files shows that orders for a certain piece of machinery were filled, respectively in 10, 12, 19, 14, 15, 18, 11 and 13 days. Use the level of significance α = -0.1 to test the

claim that on the average such orders are filled in 10.5 days. Choose the alternative hypothesis so that rejection of the null hypothesis $\mu = 10.5$ days implies that it takes longer than indicated.

Null hypothesis $\mu = 10.5$ days

Alternative hypothesis $\mu > 10.5$ days (Right tailed test)

Level of significance = .01

Test statistic $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

$X = \frac{10+12+19+14+15+18+11+13}{8} = \frac{112}{8} = 14$

$\therefore$ The mean of the sample = x = 14.

Variance of the sample = $s^2 \frac{\sum(x_i-\bar{x})^2}{n-1}$

$= \frac{(10-14)^2+(12-14)^2+(19-14)^2+(14-14)^2+(15-14)^2+(18-14)^2+(11-14)^2+(13-14)^2}{7}$

$= \frac{72}{7} = 10.3$

$t = \frac{14-10.5}{\frac{10.3}{\sqrt{8}}} = 3.08$

table value $t_\alpha = 3.00$, $t > t_\alpha$

$\therefore$ H_0 is rejected.

$\therefore$ The orders on average filled in more than 10.5 days

Test of significance for difference of means (σ unknown)

Suppose x and y are the means of two independent samples of sizes n_1 and n_2 respectively taken from a population whose mean is μ. Then $t=\frac{\bar{x}-\bar{y}-\delta}{\sigma\,\bar{x}-\bar{y}}$ where

$\sigma^2_{\bar{x}-\bar{y}} = s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$ is a random variable having the t – distribution with $n_1 + n_2 - 2$ degrees of freedom where $s^2 = \frac{\sum(x_1-\bar{x})^2+\sum(y_1-\bar{y})^2}{(n_1+n_2-2)}$

$\sum(x_i - \bar{x})^2$ = sum of the squared deviation from the mean for the first sample

$\sum(y_i - \bar{y})^2$ = sum of the squared deviations from the means for the second sample,

n_1+n_2-2 = degrees of freedom (there are n_1 -1 independent deviations from the mean in the first sample and similarly n_2-1 for 2nd sample and thus we have n_1 +n_2 -2 independent deviations from the mean to estimate the population variance)

$$\therefore t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \sqrt{\frac{n_1+n_2-2(\bar{x}_1-\bar{x}_2-\delta)}{\sum(x_1-\bar{x})^2+\sum(y_1-\bar{y})^2}} \cdot \sqrt{\frac{n_1 n_2}{n_{1+}n_2}}$$

Case (ii) when the null hypothesis is $\mu_1 = \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where $s^2 \frac{\sum(x_1-\bar{x})^2+\sum(y_1-\bar{y})^2}{(n_1+n_2-2)}$

small sample confidence interval for $\delta = \mu_1 - \mu_2$ is (100 (1-α)%) interval

$$x - y \pm t_{\alpha/2} \frac{\sum(x_1-\bar{x})^2+\sum(y_1-\bar{y})^2}{(n_1+n_2-2)} \left(\frac{1}{n_1}+\frac{1}{n_2}\right) \text{ or}$$

$$x - y \pm t_{\alpha/2}\, s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

4. Two independent samples of 8 and 7 items respectively had the following values.

Sample I	11	11	13	11	15	9	12	14
Sample II	9	11	10	13	9	8	10	-

Is the difference between the means of samples significant?

Solutions: Null hypothesis : $\mu_1 = \mu_2$

Alternative hypothesis : $\mu_1 \neq \mu_2$

Level of significance = .05 (considering)

Test statistics $t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

$\bar{X} = \frac{11+11+13+11+15+9+12+14}{8} = \frac{96}{8} = 12$

$\bar{Y} = \frac{9+11+10+13+9+8+10}{7} = \frac{70}{7} = 10$

X_1	Y_1	$X_1 - \bar{X}$	$(X_1 - \bar{X})^2$	$Y_1 - \bar{Y}$	$(Y_1 - \bar{Y})^2$
11	9	-1	1	-1	1
11	11	-1	1	1	1
13	10	1	1	0	0
11	13	-1	1	3	9
15	9	3	9	-1	1
9	8	-3	9	-2	4
12	10	0	0	0	0
14		2	4		
96	**70**		**26**		**16**

Total

$s^2 = \frac{\sum(x_1-\bar{x})^2+\sum(y_1-\bar{y})^2}{(n_1+n_2-2)} = \frac{26+16}{13} = 3.23$

$t = \frac{12-10}{\sqrt{3.23}\left(\sqrt{\frac{1}{7}+\frac{1}{8}}\right)} = \frac{2}{1.86 \text{ x } .518} = 2.159$

Conclusion : $t_{\alpha/2}$ for .05 level of significance for υ=13 is 2.16

$T < t_{\alpha/2}$ $\therefore$ H_0 is accepted

$\therefore$ The difference is not significant

5. Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results.

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether the two horses have the same running capacity (5 percent values of t for 11 degrees of freedom = 2.20)

Solution: Null hypothesis = $\mu_1 = \mu_2$

Alternative hypothesis = $\mu_1 \neq \mu_2$

Level of significance = α = .05

Test statistic : t = = $\frac{\bar{x}_1-\bar{x}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

$\bar{x}$ = mean of 1st sample

$\bar{y}$ = mean of the second sample

S = $\frac{\sum(x_1-\bar{x})^2+\sum(y_1-\bar{y})^2}{(n_1+n_2-2)}$

n_1 = first sample size =7

n_2 = second sample size = 6

x_1	y_1	x_1- $\bar{x}$	$(x_1$- $\bar{x})^2$	y_1- $\bar{y}$	$(y_1$- $\bar{y}$ $)^2$
28	29	-3.3	10.89	0.83	0.689
30	30	-1.3	1.69	1.83	3.35
32	30	0.7	0.49	1.83	3.35
33	24	1.7	2.89	-4.17	17.39
33	27	1.7	2.89	-1.17	1.145
29	29	-2.3	5.29	0.83	0.689
34		2.7	7.29		
219	169		31.43		26.613

Total

$\bar{x}$ = $\frac{219}{7}$ = 31.3

$\bar{y}$ = $\frac{169}{6}$ = 28.17

S^2 = $\frac{\sum(x_1-\bar{x})^2+\sum(y_1-\bar{y})^2}{(n_1+n_2-2)}$ = $\frac{31.43+26.613}{11}$

= $\frac{58.043}{11}$ = 5.276

t = $\frac{31.3-28.17}{2.3\left(\sqrt{\frac{1}{6}+\frac{1}{7}}\right)}$ = 2.5

Conclusion : tα = 2.2

$t > t_\alpha$ $\therefore$ Null hypothesis is rejected.

$\therefore$ Both the horses do not have the same running capacity.

6. The following are the number of sales which a sample of nine sales people of industrial chemicals in California and a sample of six sales people of industrial chemicals in Oregon made over a certain fixed period of time

California	59	68	44	71	63	46	69	54	48
Oregon	50	36	62	52	70	41			

Assuming that the populations sampled can be approximated closely with normal distributions having the same variance, test the null hypothesis $\mu_1 - \mu_2 \neq 0$ at the .01 level of significance.

Solution: Null hypothesis : $\mu_1 = \mu_2$

Alternate hypothesis : $\mu_1 - \mu_2 \neq 0$

Level of significance = .01

Test statistics $t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

Where $s^2 = \frac{\sum(x_1-\bar{x})^2+\sum(y_1-\bar{y})^2}{(n_1+n_2-2)}$

x_1	y_1	$x_1-\bar{x}$	$(x_1-\bar{x})^2$	$y_1-\bar{y}$	$(y_1-\bar{y})^2$
59	50	1	1	-1.83	3.349
68	36	10	100	-15.83	249.6
44	62	-14	196	10.17	103.43
71	52	13	169	.17	-0289
63	70	5	25	18.17	330.149
46	41	-12	144	-10.83	117.29
69		11	-121		
54		-4	16		
48		-10	100		
522	311		872		803.8469

$\bar{x} = \frac{522}{9} = 58$

$\bar{y} = \frac{311}{6} = 51.83$

$\frac{\sum(x_1-\bar{x})^2+\sum(y_1-\bar{y})^2}{(n_1+n_2-2)} = \frac{872+803.847}{13} = \frac{1675847}{13} = 128.91$

$\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \sqrt{\left(\frac{9+6}{54}\right)} = .53$

$t = \frac{58-51.83}{\sqrt{128.91 \text{ x } .53}} = 1.03$

Conclusion:

$t_{\alpha/2}$ for .01 level of significance and v = 11 is 3.01

$t < t_{\alpha/2}$ $\therefore$ H_0 is accepted

$\therefore$ There is no significant difference between the samples.

7. The average losses of workers, before and after a certain pro are given. Use 0.05 level of significance to test whether the program is effective (paired sample t-test) 40 and 35.70 and 65, 45 – 42. 120 and 116, 35 and 33, 55 and 50, 77 and 73.

Solution: Null hypothesis $\mu = 0$.

Alternative hypothesis = $\mu > 0$

Test statistic =

Before	After	di	di-$\bar{d}$	(di-$\bar{d}$)^2
40	35	5	1	1
70	65	5	1	1
45	42	3	-1	1
120	116	4	0	0
35	33	2	-2	4
55	50	5	1	1
77	73	4	0	0
		28		0

Total di

$\bar{d} = \frac{28}{7} = 4$

$S^2 = \frac{\sum(di-\bar{d})^2}{n-1} = \frac{8}{6}$

$X = d = 4$

$t = \frac{x-\mu_0}{\frac{5}{\sqrt{n}}} = \frac{4-0}{\frac{8}{6}}\sqrt{n} = 7.95$

Conclusion:

$t_\alpha = 1.94 > t_\alpha$

H_0 is rejected

$\therefore$ The program is effective

8. The incomes of a random sample of engineers in industry I are Rs. 630, 650, 680, 690, 710 and 720 per month. The incomes of a similar sample from industry II are Rs. 610, 620, 650, 660, 690, 690, 700, 710, 720 and 730. Discuss the validity of the suggestion that industry 1 pays its engineers much better than industry II.

Solution: Null hypothesis H_0, $\bar{x} = \bar{y}$

Alternative hypothesis H_1: $\bar{x} > \bar{y}$

Level of significance $\alpha = .05$

Test statistic : $n_1 = 6$

X_i	$(x_1-\bar{x})$	$(x_1-\bar{x})^2$
630	-50	2500
650	-30	900
680	0	0
690	10	100
710	30	900
720	40	1600
4080		6000

Total

$\bar{x} = \frac{4080}{6} = 680$

Second sample

y_i	$(y_1-\bar{y})$	$(y_1-\bar{y})^2$
610	-68	4624
620	-58	3364
650	-28	784
660	-18	324
690	12	144
690	12	144
700	22	484
710	32	1024
720	42	1764
730	52	2704
6780		**15360**

$n_2 = 10$

$\bar{y} = \frac{6780}{10} = 678$

$S = \sqrt{\frac{6000+15360}{(6+10-2)}} = 39.06$

$t = \frac{x-y}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{680-678}{39.06\sqrt{\frac{1}{6}+\frac{1}{10}}} = .1$

$t_{.05} = 1.76$

$t = t_\alpha$ $\therefore$ Null hypothesis is accepted.

$\therefore$ The industry I does not pay more salary than industry – II

9. Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results.

Horse A : x_i	28	30	32	33	33	29	34
Horse B: y_i	29	30	30	24	27	29	

Test whether the two horses have the same running capacity.

Solution:

Solution:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
28	-3	9	29	1	1
30	-1	1	30	2	4
32	1	1	30	2	4
33	2	4	24	-4	16
33	2	4	27	-1	1
29	-2	4	28	0	0
32	1	1	-	-	-
217		24	168		26

$$\bar{x} = \frac{217}{7} = 31, \qquad \bar{y} = \frac{168}{6} = 28$$

$$S = \sqrt{\frac{\Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2}{(n_1 + n_2 - 2)}} = \sqrt{\frac{24 + 26}{11}} = 2.13$$

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{31 - 28}{2.13\sqrt{\frac{1}{6} + \frac{1}{7}}} = 2.53$$

$$t_{.\alpha}\ for\ 11 d.f = 2.2$$

$$t > t_{.\alpha}.1\ H_0\ is\ rejected.$$

∴ There is a significant difference between the running capacities of the horses.

eg 10. The gain in weight of two random samples of rates fed on two different diets A and B are given below. Examine whether the difference in mean increases is significant.

Diet A	13	14	10	11	12	16	10	8	
Diet B	7	10	12	8	10	11	9	10	11

Solution : Null hypothesis $\mu_x = \mu_y$

Alternative hypothesis $\mu_x \neq \mu_y$

Level of significance $\alpha = .05$

Test statistic $t = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

$n_1 = 8, n_2 = 9$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
13	1.25	1.56	7	-2.78	7.73
14	2.25	5.06	10	.22	.484
10	-1.75	3.06	12	2.22	4.93
11	-.75	.506	8	-1.78	3.584
12	.25	.06	10	.22	.484
16	4.25	17.44	11	1.22	1.488
10	-1.75	3.06	9	-.78	.608
8	-3.75	14.06	10	.22	.484
			11	1.22	1.488
94		44.806	88		21.280

$$\bar{x} = \frac{94}{8} = 11.75, \qquad \bar{y} = \frac{88}{9} = 9.78$$

$$S = \sqrt{\frac{\Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2}{(n_1 + n_2 - 2)}} = \sqrt{\frac{44.806 + 2128}{8 + 9 - 2}} = 2.1$$

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{11.75 - 9.78}{2.1\sqrt{\frac{1}{8} + \frac{1}{9}}} = 1.9$$

$t_{.025}(15\ d.f) = 2.13 \qquad t < t_\alpha \ \therefore Null\ hypothesis\ accepted.$

$\therefore$ There is no significant difference between the means.

eg 11. Two independent groups of 10 children were tested to find how many digits they could repeat from memory after hearing them. The results are as follows

Group A	8	6	5	7	6	8	7	4	5	6
Group B	10	6	7	8	6	9	7	6	7	7

Is the difference between the mean scores of the two groups significant?

Solution : Null hypothesis $\mu_1 = \mu_2$

Alternative hypothesis $\mu_1 \neq \mu_2$

Level of significance $\alpha = .05$

Test statistic $t = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
8	1.8	3.24	10	2.7	7.29
6	-.2	.04	6	-1.3	1.68
5	-1.2	1.44	7	-.3	.09
7	.8	.64	8	.7	.49
6	-.2	.04	6	-1.3	1.69
8	1.8	3.24	9	1.7	2.89
7	.8	.64	7	-.3	.09
4	-2.2	4.84	6	-1.3	1.69
5	-1.2	1.44	7	-.3	.09
6	-.2	.04	7	-.3	.09
62		15.60	73		16.10

$$\bar{x} = \frac{62}{10} = 6.2, \qquad \bar{y} = \frac{73}{10} = 7.3$$

$$S = \sqrt{\frac{\Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2}{(n_1 + n_2 - 2)}} = \sqrt{\frac{15.6 + 16.1}{18}} = 1.32$$

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{6.2 - 7.3}{1.32\sqrt{\frac{1}{5}}} = 1.86$$

$t_\alpha(18\ d.f) = 2.1 \quad t < t_\alpha \ \therefore Null\ hypothesis\ accepted.$

$\therefore$ There is no significant difference between the means.

eg 12. The table gives the biological value of protein from 6 cow's milk and 6 buffalo's milk. Examine whether the differences are significant.

Cow's milk	1.8	2.0	1.9	1.6	1.8	1.5
Buffalo's milk	2	1.8	1.8	2.0	2.1	1.9

Solution : Null hypothesis $\mu_1 = \mu_2$

Alternative hypothesis $\mu_1 \neq \mu_2$

Level of significance $\alpha = .05$

Test statistic $t = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

Cow's milk			**Buffalo's milk**		
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1.8	.03	.0009	2.0	.07	.0049
2.0	.23	.0529	1.8	-.13	.0169
1.9	.13	.0169	1.8	-.13	.0169
1.6	-.17	.0289	2.0	.07	.0049
1.8	.03	.0009	2.1	-.17	.0289
1.5	-.27	.0729	1.9	-.03	.0009
Total 10.6		.1734	11.6		.0734

$\bar{x} = \frac{10.6}{6} = 1.77, \qquad \bar{y} = \frac{11.6}{6} = 1.93$

$S^2 = \frac{.1734+.0734}{10} = \frac{.2468}{10} = .0247$

$S = .155$

$t = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{1.77-1.93}{.155\sqrt{\frac{1}{6}+\frac{1}{6}}} = -1.1$

Conclusion : $t_\alpha(10\ d.f) = 2.23 |t| < t_\alpha$

$\therefore$ *Null hypothesis accepted.*

$\therefore$ There is no significant difference between the means.

eg 13. For a random sample of 10 pigs, fed on a diet A, the increase in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 lbs, for another random sample of 12 pigs fed on diet B. The increase in the same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17, lbs. Find if the two samples are significantly different regarding the effect of diet.

Solution : Null hypothesis $\mu_1 = \mu_2$

Alternative hypothesis $\mu_1 \neq \mu_2$

Level of significance $\alpha = .05$

Test Statistic $t = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
10	-2	4	7	-8	64
6	-6	36	13	-2	4
16	4	16	22	7	49
17	5	25	15	0	0
13	1	1	12	-3	9
12	0	0	14	-1	1
8	-4	16	18	3	9
14	2	4	8	-7	49
			21	6	36
15	3	9	23	8	64
9	-3	9	10	-5	25
			17	2	4
Total 120		120	180		314

$\bar{x} = \frac{120}{10} = 12,$ $\bar{y} = \frac{180}{12} = 15$

$$S = \sqrt{\frac{\Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2}{(\mathrm{n}_1 + \mathrm{n}_2 - 2)}} = \sqrt{\frac{120 + 314}{20}} = 4.6$$

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{12 - 15}{4.6\sqrt{\frac{1}{10} + \frac{1}{12}}} = -1.5$$

Conclusion : $t_\alpha(20\ d.f) = 2.09$

$$t < t_\alpha$$

$\therefore$ *Null hypothesis accepted*.

$\therefore$ There is no significal difference between the two means.

Unit-V

Chi-square and Analysis of Variance – chi-square as test of independence, chi-square as a test of goodness of fit, analysis of variance, Inferences about a population variances.

Regression and Correlation – Simple Regression – Estimation using regression line, correlation analysis, making inferences about the population parameters, limitations, errors and caveats in regression and correlation analysis, multiple regression and correlation analysis, Finding multiple regression equations and making inferences about population parameters.

CHI (χ^2) SQUARE TEST

In the earlier sections, problems of estimation and hypothesis testing were solved by means of the Normal Distribution or the t-distribution. Normal Distribution either described the sampling distribution or approximated the Binomial Distribution for large samples. The t-distribution was used for small samples.

Where the Binomial distribution (approximated by the Normal) was used, the outcome could be classified as a success or a failure. There are several problems where the experimental outcome required more than two categories of classification. A simple example would be rolling of a single die.

Furthermore, in the problems to the estimation or significance testing of the population parameter a, the standard deviation were not dealt with.

A remarkable feature of these two seemingly unrelated classes of problems is that their solution depends upon the same probability distribution. This distribution is called the Chi-square Distribution. It can even be employed for handling certain correlation problems. Obviously, it is a versatile distribution.

The important applications of χ^2 test are

(i) test for goodness of fit

(ii) test for independence of attributes.

(iii) test for a specified S.D.

χ^2 test is applied to test the association between attributes when the sample data is presented in the form of a contingency table with any number of rows, *r,* and columns *c.*

The null and alternative hypotheses are set as follows:

H_0 : No association exists between the attributes.

H_1 : An association exists between the attributes.

The following steps are required to perform the test of hypothesis.

Step 1. An expected frequency, E, corresponding to each cell in the table is found by using

$$E = \frac{R*C}{N}$$

where R = row total,

C = column total

and N = Total frequency.

Step 2. Based upon the observed values and corresponding expected frequencies, the statistic χ^2 is found with the help of

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

The characteristics of this distribution are completely defined by the number of degrees of freedom, d. f., which is given by

d. f. = (r-1)(c-1),

where r = number of rows, and c = number of columns

Step 3. Corresponding to a chosen level of significance, $\propto$, the critical value of $\chi^2_{\propto}$ corresponding to the number of d. f. is found in the table.

Step 4. The computed and the table values of $\boldsymbol{\chi^2}$ are compared

H_o is accepted when the computed value is < the critical value;

H_o is rejected (i.e. H_I is accepted) when the computed value $\geq$ the critical value.

As an example, consider the experiment of rolling a 6 faced die. It is to be ascertained if the die is "honest". The results of performing the experiment are tabulated in the cells below. Also, the expected frequencies on the assumption that the die is honest are given. Here total throws, n=60.

	1	2	3	4	5	6
Observed	1	8	6	1	9	1
Expected(	1	1	1	1	1	1

The general method of testing the compatibility is based on the extent to which the observed frequencies are close to the expected frequencies.

The value of $\boldsymbol{\chi^2}$, in this case, is given by

$$\boldsymbol{\chi^2} = \frac{(12-10)2}{10} + \frac{(8-10)2}{10} + \frac{(6-10)2}{10} + \frac{(10-10)2}{10} + \frac{(9-10)2}{10} + \frac{(15-10)2}{10}$$

$$= 5.0$$

For perfect agreement amongst the observed frequencies and the expected frequencies the value of $\boldsymbol{\chi^2}$ should obviously be zero. The large the value of $\boldsymbol{\chi^2}$,the proper the agreement between the observed and expected frequencies. If the disagreement is significant to lead to the conclusion that the die is not honest, $\boldsymbol{\chi^2}$ distribution provides the answer. Since there are only a limited number of classe

Example. A certain drug ins claimed to be effective in curing colds. In n experiment on 164 people with cold, half of them were given the drug and half of them given sugar pills. The patients' reactions to the treatment are recorded in the following table Test the hypothesis that the drug is not better than sugar pills for curing colds.

	Helped	Harmed	No .effect
Drug	52	10	20
Sugar pills	44	12	26

Solution

Expected frequency corresponding to A = 82x96/164 = 48.

Other expected frequencies have been computed similarly, using

grand total= Expected frequency

Helped	Harmed	No effect	Total
52 A=48	10 B-11	20 C=23	82
44 D=48	12 E;=11	26 F-23	82
96	22	46	164

x2 may now be calculated as follows:

□2 = (52-48)2/48 +(10-11)2/11 + (20-23)2/23 + (44-48)2/48 + (12-11)2/11 + (26-23)2/23

= 1.622

Now the degrees 'of freedom are calculated as (r-1)(c-1), where r=number of rows; and number of columns.

In this problem, therefore, d f. = (2-1)(3-1)=2. From the table, for 2 d f.

Since the computed value of X2 is less than the critical value at 5% significance level, it is to be concluded that the experiment does not provide evidence that the drug is better than sugar pill.

χ2 TEST FOR GOODNESS OF FIT

Earlier method of fitting Binomial, Poisson and Normal Distribution to a given data was ex¬plained. But it remains to settle if the distribution provides sufficiently good fit, that is if the discrep¬ancies between the actual frequencies and the theoretical frequencies based on the distribution being fitted are too great to dismiss the distribution-as the underlying one.

The following example should bring out clearly the use of the x2-distribution in this regard :-

Example: In 1,000 extensive sets of trials for an event of small probability, the frequencies o of the number x_i of successes proved to be

x_i	0	1	2	3	4	5	6	7
o	305	365	210	80	28	9	9	1

Find the mean and variance of the successes. Discuss the possibility of frequencies following the Poisson Distribution. Compute the expected frequencies by this distribution. Also, test the goodness of fit by the Chi-square test.

x_r	O_1	x_1O_1		O_1X^2
0	305	0	0	0
1	365	365	1	365
2	210	420	4	840
3	80	240	9	720
4	28	112	16	448
5	9	45	25	225
6	2	12	36	72
7	**1**	7	49	49
	1,000	1,201		2,179

Mean = 1201/1000 = 12.01 1.20

Variance = 2719/1000-(1.2)2 = 1.279 1.28

Since the mean is approximately equal to the variance there are reasons to believe that the successes conform to the Poisson Distribution.

The general expression for the Poisson distribution is given by, f(x) = mx/x!, e-x

The theoretical frequencies 0,1,2…..7 successes would be obtained from the following series.

$$100\left[e^{-1.2}\left\{1\ \frac{1.2}{1!}+\frac{1.2}{2!}+\frac{1.2}{3!}+\frac{1.2}{4!}+\frac{1.2}{5!}+\frac{1.2}{6!}+\frac{1.2}{7!}\right\}\right]$$

These are tabulated below against the actual frequencies.

x_i	0	1	2	3	4	5	6	7
O_i	305	365	210	80	28	9	2	1
E_i	301.2	361.4	216.8	86.7	26.6	6.2	1.2	0.2

Since the expected or the theoretical frequencies in the last 2 classes are than 5, these are amalgamated along with x_i= to x_1>5, as below:

x_i	0	1	2	3	4	5
	305	365	210	80	28	12
E_i	301.2	361.4	216.8	86.7	26.0	7.6

$$\chi^2 = \frac{(305-301.2)^2}{301.2}+\frac{(365-361.4)^2}{351.4}+\frac{(210-216.8)^2}{216.8}+\frac{(85-86.7)^2}{86.7}+\frac{(28-26)^2}{26}+\frac{(12-7.6)^2}{7.6}=3.5$$

d.f = 6-1-1=4

$x^2_{0.05}$ =11.688

The computed value does not fall in the critical region. Therefore, the assumption of Poisson-distributed number of successes is valid.

A note on Degrees of Freedom

In a 2x2 contingency table, if the totals are known, just one of the frequencies need to be ascertained. Others can be ascertained by proper subtraction from the totals.

	A	B	Total
a			(known)
13			(known)
Total	(known)	(known)	

Generally, in a contingency table, d.o.f. = (c-1)(r-1)

For the poisson distribution d.o.f. = Number of classes -1 -the number of parameters estimated from data = Number of classes -2, because the mean is only statistic that need to be computed to describe the distribution completely.

Similarly, d.o.f. for normal distribution is number of classes -2, since mean and standard deviation both need to be estimated from the data. And for the binomial distribution, d.o.f. = Number of classes - 1.

ANALYSIS OF VARIANCE

This is one of the. most elegant and versatile statistical technique and finds wide application n determining whether or not the means of more than two populations are equal. Basically it is a procedure by which the variation is embodied in the data. It is the technique to test the homogeneity of several means.

The analysis of variance is a procedure which separates the variation ascribable to one set of causes from the variation ascribable to other set of causes. For example, four varieties A, B, C, D of wheat are sown in plots and their yield per acre was recorded. We want to test the null hypothesis that the four varieties of wheat produce an equal yield. However it can be done by taking all the possible

pairs of means and testing the significance of their difference simultaneously. The variation in yield may be due to various factors, like varieties of wheat, fertility of soil, use of different kind of fertilizers etc. If we are interested to whether the variation in yield is due to varieties of wheat, use of types of fertilisers or in both. Analysis of variance is a method to estimate the contribution made by each factor to the total variation. The total variation is split into the following two parts : (i) variation between the samples and (ii) variation within the samples.

UNDERLYING ASSUMPTIONS

1. Each of the samples is a simple random sample.
2. Populations from which the samples are selected are normally distributed.
3. Each of the populations has the same variance.
4. Each of the sample is independent of the other samples. If, however, the sample sizes are large

The technique of analysis of variance is referred to as ANOVA. A table showing the source of variation, the sum of squares, degrees of freedom, mean square (variance), and the formula for the F ratio is known as ANOVA table.

Let us adopt a standard arrangement of the data. A sample observation has two subscripts. The subscript *i* denotes the row or sample observation, while the subscript *j denotes* the column or population from which the observation came.

***F-TEST** :*

This is another non-parametric test of significance based on variance named after its profounder R.A. Fisher. Its application is there is case of inference when two or more samples are involved and the test in made on the basis of variances. The variances of random samples drawn from a normal universe have a distribution that is skewed positively; therefore we have a separate distribution called F-distribution to test the significance based on variances.

The various sums of squares involved in this are as follows:

(i) Sum of the squares of variations amongst the columns (SSC): It is the sum of the squares of deviation between column or group and means the grand mean. We can indicate it as

$$ri\Sigma(xi - \bar{x})2$$

where x_i is the mean of the j-th sample, *r* is the number of rows (size of each sample), x is the mean of the sample (column) means.

then this is divided by c-1 (the number of columns minus one) or the degrees of freedom amongst columns, we have, mean of the sum of the squares of deviations between mean of the columns and the grand mean designated as MSC. This is called the variance amongst columns and indicates the degree of explained variance due to sampling variations. It, in other words indicates something more than a chance variation which is probably due to unlink populations.

(ii) **Sum of the square of variations within columns (SSE):** It is the sum of the squares of variations between individual items and the column means. It is given by SSE = $(x_o\text{-}x_i)^2$, where is i-th observation in the j-th sample and x_i, is the mean of the j th column. This, when divided by c(r-1), where *c* is the no. of columns and *r* is the no. of rows, we have MSE, the means of the square of column errors. This is also called unexplained variance because it indicates only the chance variation which cannot be explained in items of variation in the population.

(iii) **Total sum of squares of variations (SST):** The total sum of squares is given by SST = $\sum x_i x_j$- C, where C=correction error = T^2/rc, T being the grand total of the values in all the samples. r, c having their meaning
introduced earlier. The total sums of squares refers to the total of SSC and SSE. It is the sum of square of observations between the individual values and the grand mean. We can write it as

SST = SSC + SSE

This when divided by *n-1* or the degrees of freedom, gives the total variance comprising both the explained and the unexplained variance.

F-value : The F-value is the ratio of unexplaijed variance (MSC) to the explained variance (MSE). We can indicate it as

F = MSC / MSE

Following is ANOVA table for one-way classification for equal sample size

Sources of variation	Sum of squares	Degrees of freedom	Mean sum of square	*F*
Between samples (column means)	SSC	d.f.=c-1	MSC=SSC/c-1	F=MSC/MSE
Within samples (errors)	SSE	d.f.=c(r-1)	MSE=SSE/c(r-1)	
Total	SST	cr-1		

If the sample sizes are unequal, c(r-1) is to be replaced by N-c, and cr-1 by N-1, N being the total number of observations in all the samples.

The rationale for the test is simple if the variance amongst the columns or groups (MSC) is equal to what is there within the group (MSE); the F-value will be equal to one showing no extra

variation in columns than what is there within the columns, and therefore no variance worth the notice. If however, the variance amongst columns is more than that within the columns, the derived F value has to be compared with the table value.

If the derived value is more than the table value the different is significant and the null hypothesis is rejected. If the derived value is less than the table value the

difference is not significant and the null hypothesis is accepted. The actual application will be clear from the following illustrations.

Example: The three samples below have been obtained from normal populations with equal variance. Test the hypothesis at 5% level that the population means are equal

8	7	12
10	5	9
7	10	13
14	9	12
11	9	14

(The table value of F at 5% level of significance for v_i =2, and v_2 =12 is 3.88)

Solution

We set up the null Hypothesis : $H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : At least two of the population means are unequal.

	8	7	12	27
	10	5	9	24
	7	10	13	30
	14	9	12	35
	11	9	14	34
Group total	50	40	60	T=150
				Grand mean
Group mean $\bar{x}_i$	10	8	12	x = 10

The sum of square for columns = SSC = $rj\Sigma(x_j-x)^2$, where x_j is the mean of the jth sample, r is the number of rows, and $\bar{x}$ is the mean of he sample (columns) means,

$$SSC = 5\{[10-10]^2 + [8-10]^2 + [12-10]^2\} = 5 \times 8 = 40$$

The sum of squares for columns MSC = SSC/c-1. where *c* stands for number of columns. MSC = 40/2 = 20

The sum of squares for the error = SSE = $\Sigma_i\Sigma_i(x_{ii}-\bar{x}_i)^2$, where x_{1i} is the j^{th} observation in the j^{th} sample and xi is the mean of the column.

SSE = $(8-10)^2+(1\ 0-10)^2+(7-10)^2+(1\ 4-10)^2+(11-10)^2+(7-8)^2+(5-8)^2+(1\ 0-8)^2+(9-8)^2+(9-8)^2+(12-12)^2$ -(9-12)2+(13-12)2+(1 2-12)2+(14-12)2

= 4+0+9+16+1+1+9+4+1+1+0+9+1+0+4 = 60

Variance within columns = MSE = SSE/c(r-1) = 60/3x4 = 5.

Total sum of squares of variations = SST = SSC+SSE = 40+60 = 100

d. f = c-1 = 2; d.f. = c(r-1) = 3x4 = 12

F = MSC/MSE = 20/5 = 4

Source of Variation	Sum of squares	Degrees of freedom	Mean square	F
Between samples (column means)	SSC=40	d. f = 2	MSC=20	F=MSC/MSE =4
To t a l	SST=100	cr-1=14		

At 5% level the table value of F for v_1=2 and v_2=12 is given to be 3.88 but the computed value of F is greater than this table value. We, therefore, reject H_o and conclude that the population means are not equal.

Alternative approach for computation of SSC, SST, SSE.

Correction factor, $C=T^2/rc$, where T is the grand total of the values in all of the sample, r is the number of rows, and e is the number of column.

= 150x150/5x3 = 1500

Now SSC = $\Sigma T^2/r$-C, where T stands for th total of the jth columns.

SSC = $\frac{50^2+40^2+60^2}{5}$ -1500 = 1540 – 1500 = 40

SST = $\Sigma_i T_j x^2_{ij}$-C=(8^2+10^2+7^2+14^2+11^2+7^2+5^2+10^2+9^2+12^2+9^2+13^2+12^2+14^2)-1500

= 1600-1500=100

SSE = SST-SSC=100-40=60

Analysis of Variance in manifold classification : In manifold classification there are two or more characteristics which are considered. The table dealing with data on such grouping has a number of columns and rows. The analysis in that case will get extended to include the sum of squares between rows, which was not there is one-way classification. See the table below:

Source of variation	Sum of Squares	d.o.f.	Mean Squares Ratio	Variance
Between	SSC	(c-1)	SSC/c-1=MSC	F_C = MSC/MSE
Between Rows	SSR	(r-1)	SSR/r-1=MSR	FR= MSR/MSE
Error	SSE	(c-1) (r-1)	SSE/(r-1)(c-l) = MSE	
Total	SST	cr-I		

Example: A farmer applies three types of fertilizers on 4 separate plots. The figure on yield per acre are tabulated below:

Total	Yield				
Fertilizers/Plots	A	B	C	D	T
Nitrogen	6	4	8	6	24
Potash	7	6	6	9	28
Phosphates	8	5	10	9	32
Total	21	15	24	24	84

Find out if the plots are materially different in fertility as also if the three fertilizers make any material difference in yields.

Solution: Let us first determine the correction factor.

$$C = T2/cr = (84)2/12 = 588.$$

The sum of squares between column is SSC = $(\Sigma x_i)^2/n_j$- C

$$=\left[\frac{21^2}{3}+\frac{15^2}{3}+\frac{24^2}{3}+\frac{24^2}{3}\right]\text{-}588$$

$= \frac{1818}{3}$ - 588 = 606 -588 = 18

The sum of squares between rows is, SSR = $\sum_{i=1}^{r} T_i^2 / C^{-c}$

$= \left[\frac{24^2}{3} + \frac{28^2}{3} + \frac{32^2}{3}\right]$ -588 = $\left[\frac{576}{4} + \frac{784}{4} + \frac{1024}{4}\right]$ -588

= 2384/4-588 = 596-588 =8.

Total sum of square is, SST = $\sum_{i-1}^{r} \sum_{j=1}^{C} x_{ij}^2 - C$

=[6^2+7^2+8^2+4^2+6^2+5^2+8^2+6^2+10^2+6^2+9^2+9^2]-588

=[36+49+64+16+36+25+64+36+100+36+81+81]-588

= 624-588 =36

The error sum of squares is, SSE=SST-(SSR+SSC)=36-(8+18)=10.

The table for Analysis of Variance :

Source of Variation	Sum of Squares		Degree of freedom	Mean Sqares	F-Rations
Between plots or	18	4-1=3	MSC=18/(4-1)=6	F_c=6/1.667=3.6	Columns (SSC)
Between fertilizer	8	3-1=2	MSR=8/(3-1)=4	F_R=4/1.667=2.4	Or rows (SSR)
Error (SSE)	10	3x2=6	MSE=10/6=1.667		
Total	36	cr-1=12-1=11			

The F –value for (3,6) degrees of freedom is 4.76 and for (2, 6) degrees of freedom is 5.14 both at 5% level of significance. The computed F values being lower, they do not show any significant difference, and whatever difference exists is due to sampling error.

CORRELATION AND REGRESSION

INTRODUCTION

In many business situations we have to deal with two or more variables. Specially, in the analysis and interpretation of data we have to take into account the relationship between two or more variables. For example, we may have to find out relationship between demand and price, output and rainfall, volume of sales and expenditure on advertisement, etc. For study of such relationships the two important statistical methods used are correlation and regression.

These methods are also helpful in forecasting figures for the future. For example, a company planning next year's production may be interested in the forecast of sales for that year. If the marketing manager knows the sales have a relationship with advertising expenditure and few other variables such as public expenditure, national income, etc., he will be able to predict the value of sales with the help of known relationship of the sales with these variables provided the value of all these variables is known. Similarly, a cost accountant can estimate the cost of a product if there are established relationship between the cost and the price of inputs such as labour, material, sales promotion expenditure, etc. In statistics we find these relationships by the methods of correlation and regression.

CORRELATION

Correlation refers to the statistical tool for measuring the degree of relationship that exists between two or more variables. Such a measure determines how well a mathematical law explains the relationship between the variables. It denotes the inter dependence between the variables. If the change in one variable affects a change in other variable then variables are said to be correlated. The study of correlation is specially important in social science, educational research, policy making and arriving at decisions, etc. While studying problems, when we know little about cause and effect, a knowledge of the relationship between cause and

effect is useful in drawing a conclusion if there is high degree of correlation. The amount of correlation in a sample is measured by the sample coefficient of correlation, and is denoted by r.

NATURE OF CORRELATION

When high value of one variable are associated with high values of the other variable, they are said to be directly or positively correlated. When high values of one tend to accompany low values of the other, they are inversely or negatively correlated. Two variables may be highly, slightly, or moderately correlated.

PROPERTIES OF CORRELATION COEFFICIENT

1. It is independent of the choice of both origin and scale observation.
2. It is a pure number and is independent of the units of measurement.
3. It lies between -1 to +1.

VARIOUS METHODS TO MEASURE CORRELATION COEFFICIENT

The following are the important methods to measure correlation coefficient.

(i) Scatter Diagram

(ii) Graphic method

(iii) Karl Pearsons Coefficient of Correlation

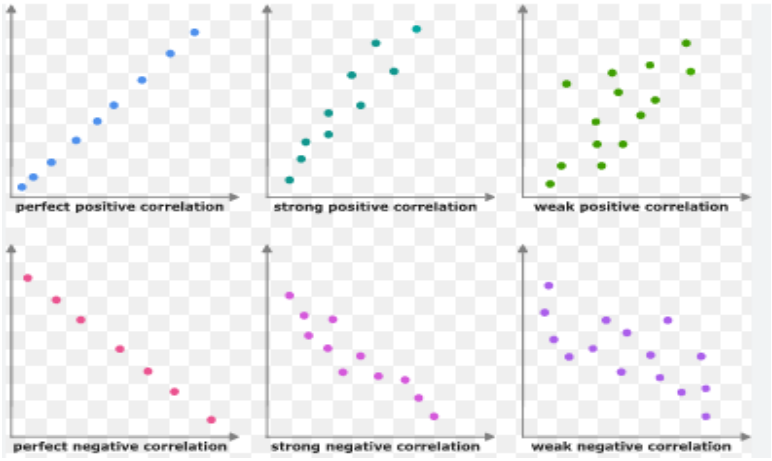
(iv) Rank Correlation Method

(v) Concurrent Deviation Method

Important methods we will discuss in the following sections.

(i) Scatter Diagram : It is the diagrammatic representation of the two variables. In this method , the two variables are shown on the two axes, the horizontal axis for one variable and the vertical axis for the other variable. The pair of values are shown by points in the space. The scatter of these points as also the direction of

scatter reveals the nature and degree of correlation between the two variables. The following are some models of scatter diagrams revealing different types of correlation between two variables.



The adjoining diagrams deal with 5 such situations. You will find that relationships shown by these diagrams are not very precise. The method is only a very rough measure of correlation where the exact magnitude cannot be known.

(ii**) GRAPHIC METHOD :** It is another method of ascertaining of correlation between the two variables. The individual values of the two variables, X and Y are plotted on a graph paper. We, thus obtain two curves, one for the x variable and the other for y variable. These two curves form the basic of comparison whether the two variables are correlated or not. If both the curves moves in the same direction , ie., parallel to each other either upward or downward; the correlation s said to be positive. If they are moving in opposite direction then s, then the correlation is said to be negative. Such type of graphical representations common is common in ecological, environmental and economics and business ares.

(iii) KARL PEARSON'S COEFFICIENT OF CORRELATION

The karlpearson's coefficient of correlation between two variables X and Y are denoted by r(X,Y) or simply r_{xy} and is given by

$$r = \frac{\sum(x-\overline{X})(y-\overline{Y})}{\sqrt{(x-\overline{X})^2(y-\overline{Y})^2}}$$

where x is the independent variable and y is dependent variable. Where as $\bar{x}$ and $\bar{y}$ are mean of the independent variable and dependent sets of variables, respectively. However , (x – $\overline{X}$) and (y – $\overline{Y}$) are the deviations from the mean, then d_x and d_y will stand for (x – $\bar{X}$) and (y - $\bar{Y}$). Now the formula becomes as follows :

$$r_{xy} = \frac{\sum d_x d_y}{\sqrt{d_x^2 d_y^2}}$$

The simplification of the components of the formula are as follows

$$\sum d_x d_y = \sum xy - \frac{\sum x \sum y}{n}$$

$$\sum d_x^2 = \sum xy - (\sum x)^2 / n$$

$$\sum d_y^2 = \sum xy - (\sum y)^2 / n$$

Where $\sum d_x d_y$ can be either the positive or negative, depending upon whether the correlation is a positive or negative. However, the correlation coefficient formula can be rewritten as follows :

$$r = \frac{\sum xy - \sum x \sum y / n}{\sqrt{[\sum x^2 - (\sum x)^2 / n]} \sqrt{[\sum y^2 - (\sum y)^2 / n}}$$

Correlation always lies between -1 and +1.

We will study the use of this method in case of individual distribution and the group distribution . first we take case of an individual distribution.

Example. Find the correlation coefficient between sales and advertising expenditure from the following data.

Sales (Rs. lakhs)	65	66	67	67	68	69	70	72
Advertising expenditure (Rs. 000)	67	78	65	68	72	72	69	71

Solution:

Sales (Rs.lakhs) X	Advertising expenditure (Rs. 000) Y	Deviations from the mean $d_x = X - \bar{X}$	Square of deviations Expenditure d_x^2	Deviation from the mean $d_y = Y - \bar{Y}$	Square of deviations d_y^2	Product of deviations $d_x d_y$
65	67	-3	9	-2	4	6
66	68	-0	4	-1	1	2
67	65	-1	1	-4	16	4
67	68	-1	1	-1	1	1
68	72	0	0	+3	9	0
69	72	+1	1	+3	9	3
70	69	+2	4	0	0	0
72	71	+4	16	+2	4	8
544	550		36		44	24

$\bar{X} = \sum X/n = 544/8 = 68$; and $\bar{Y} = \sum Y/n = 550/10 = 55$

As per Karl Pearson Coefficient of Correlation =

$$r_{xy} = \frac{\sum d_x d_y}{\sqrt{d_x^2 d_y^2}} = \frac{24}{\sqrt{36x44}} = 0.603$$

Thus , there is a fair degree of positive correlation between the volume of sales and the advertising expenditure.

.When deviation are taken from the assumed average and not the acutal average ,the corrected formula is as follows :

$$r = \frac{n\sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{[\{N\sum d_x^2 - [(\sum d_x)^2\}\{N\sum d_x^2 - \sum d_y{}^2\}]}}$$

$$r = \frac{\sum d_x d_y - [(\sum d_x)(\sum d_y)/2]}{\sqrt{[\{\sum d_x^2 - [(\sum d_x)^2/n\}\{\sum d_y^2 - \sum d_y{}^2/n\}]}}$$

let us take an example to illustrate the use of this formula.

Example. Calculate coefficient of correlation between X and Y:

X	78	89	97	69	59	79	68	61
Y	125	137	156	112	107	136	123	108

Solution

X	Y	Dev. From as.av.(69) dx	Dev. From as.av.(112) dy	Dx^2	Dy^2	dxdy
78	125	+9	+13	81	169	+117
89	137	+20	+25	400	625	+500
97	156	+28	+44	784	1936	+1232
69	112	0	0	0	0	0
59	107	-10	-5	100	25	+50
79	136	+10	+24	100	576	+240
68	123	-1	+11	1	121	-11
61	108	-8	-4	64	16	+32
n = 8	n =8	+43	+108	+1530	+3468	+2160

Substituting the values in the above formula , we have

$$r = \frac{8x2160-48x108}{\sqrt{[8x1530-(48)^2][8x3468-(108)^2]}}$$

$$r = \frac{17280-5184}{\sqrt{(12240-2304)(27744-11664)}}$$

$$r = \frac{12096}{\sqrt{9936x16080}} = \frac{252}{\sqrt{207x335}}$$

=0.99

There is thus a significant positive correlation between two variables.

(iv) Rank Method: The method is based on the rank or the order and not the magnitude of the variable as in the earlier method. As such it is more suitable when the variables can better be arranged: e.g., in the case of intelligence or beauty. The ranks range from 1 to n. Edward Spearman has provided the following rule for calculating rank correlation coefficient.

$$r = 1 - \frac{\sum 6D^2}{n(n^2-1)}$$

$$r = 1 - \frac{\sum 6D^2}{n^3-n}$$

Example : calculate the coefficient of correlation from the following data by the method of rank correlations:

X	75	88	95	70	60	80	81	50
Y	120	134	150	115	110	140	142	100

Solution :

X	Rank	Y	Rank	Rank difference D	Square of rank difference D^2
75	5	120	5	0	0
88	2	134	4	2	4
95	1	150	1	0	0
70	6	115	6	0	0
60	7	110	7	0	0
80	4	140	3	1	1
81	3	142	2	1	1
50	8	100	8	0	0
					6

Coefficient of correlation $r = 1 - \frac{\sum 6D^2}{n(n^2-1)} = 1 - [6x6 - 8(64-1)] = 1 - 36/504 = 0.9$

Application of correlation to some special cases :

1) In the case of time series the correlation between two variables should be calculated in the usual manner after calculating the trend values in the concerned variables.

2) If the two series are associated by a lag or a lead of some period then the correlation analysis should be conducted only after adjusting the values so that the adjusted pair of values are the truly related values. For example, the price of an article is influenced by supply by the price of the current month is influenced by the supply of the previous month. The correlation should be calculated in usual way taking the supply of the previous month with the price of the current month for calculation of coefficient of correlation.

3) Sometimes the data given are such that it becomes difficult to decide as to in which of the two variables the coefficient of correlation has to be calculated.

Probable Error : The Probable error of the coefficient of correlation is calculated as follows:

$$P.E.= 0.6745\ (1-r^2)/\sqrt{N},$$

The probable error or the standard error are used for interpreting the coefficient of correlation. In regard to the probable error the interpretation is done as follows:

1. If the value of r is less than the probable error there is no evidence of correlation, ie., the value of r is not at all significant.
2. If the value of r is more than six times the probable error the existence of correlation is practically certain i.e., the value of r is significant.
3. By adding and subtracting the value of probable error from the coefficient of correlate on we get respectively the upper and lower limits within which coefficient of correlation in the population can be expected to lie. Symbolically

Correlation of the population = r ± P.E.

The use of standard error is also for the same purpose. Rather to ensure better realise and therefore it is in more use than the P.E

Example. If r=0.6 and N-64, find out the P.E. of the coefficient of correlation and determine the limits for the measure for population.

P.E. = 0.6745 - $(1-r^2/\sqrt{N})$, where N is the number of items

When r = 0.6 and N-64,

P.E. = 0.6745[1-(0.6)2/8] = 0.6745x0.64/8 = 0.54

Limits of population are 0.6 + 0.054 i.e., 0.546 to 0.654

Coefficient of determination : It is square of the coefficient of correlation

Coefficient of correlation	Coefficient of determination
r	r^2
0.90	0.81
0.80	0.64
0.70	0.49
0.60	0.36
0.50	0.25

Coefficient of determination is preferred to coefficient of correlation because it explains the production of variation in the dependent variable which is explained by a change in the independent variable. We shall consider the other significance of this measure after we have studied regression analysis.

REGRESSION ANALYSIS :

By regression we mean average relationship between two or more variables. One of these variable is called the dependent or the explained variable and the other variable, the independent or the explaining variable. If the explaining variables are two or more, then it will be called the multiple regression analysis.

Regression takes its name from studies made by Sir Francis Galton. He compared the heights of persons to the heights of their parents. His major conclusion was that the off-springs of unusually tall persons tend to be shorter than their parents while children of usually short parents tend to be

shorter than their parents. In a sense, the successive generations of off-springs from tall persons "regress" downward toward the height of the population, while the reverse is true originally for short families. But the distribution of heights for the total population continues to have the same variability from generation to generation.

DISTINCTION BETWEEN REGRESSION ANALYSIS AND CORRELATION

In case of regression analysis there is a functional relationship between Y and X such that there is only one value of X. One of the variables is identified as a dependent variable and other(s) as independent variable(s). The expression is derived for the purpose of predicting values of a dependent variable on the basis of independent variable.

There arise certain situations in the analysis of business problems in which an expression of the degree of association between variables is desired, and the ability to predict values of the dependent variable may not be of any particular value. In this type of situation the correlation analysis is applicable. In correlation analysis Y may be a function of X and X may be a function of Y. The analysis is not a functional dependence of any one on the other but of mutual variation and association.

TYPES OF REGRESSION

Broadly we can divide regression analysis into two classes, simple and multiple. In case of the former, the relationship is between two variables only; one of them is the explaining or the independent variable. In the case of multiple regression there are more than one explaining variable, e.g., the output may be related with the application of fertilisers, rainfall and the number of ploughings done. Each of

these regression analysis can be further divided into (a) linear; (b) curve-linear line.

In the linear relation the dependent or the explained variable varies at a constant rate with a given change in the independent or the explaining variable. The constant rate of change can be in absolute terms or in terms of rate or percentage. A constant rate of change on a log scale also yields a straight line.

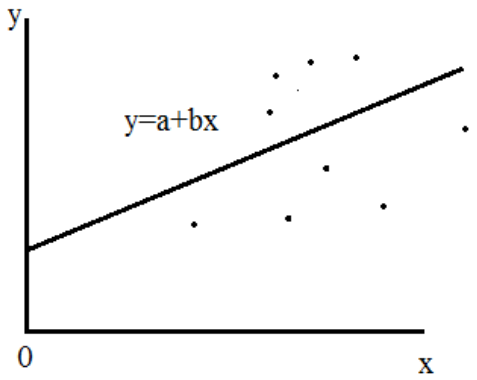
In the curve-linear relationship the explained variable changes at varying rate with a given change in the explaining variable. For example, the output of wheat increases rapidly with the application of the initial dose of fertilizer; thereafter it increases at a falling rate. The relationship in such case, when shown on a graph will yield a curve.

Whatever may be nature of relationship i.e., simple or multiple, linear or curve-linear, geometrically it can be represented on a graph and algebraically through an equation. We have therefore, to arrive at either the line or an equation of the line showing such a relationship. This is arrived at one the basis of the available data by the use of certain methods Regression analysis will become very complicated and tedious for manual calculation if we take into account curve-linear relation or when we take a large number of variables. For illustrative purposes we will confine only to simple linear relations. It is the feeling of experts that most business phenomena can be explained through linear relationship and by confining to only small number of variables. Particularly, when we adopt the graphic method, the relationship between only two variables is often considered on account of the facility in drawing two dimensional graph.

(i)Graphic Method: Under this method the two variables are plotted on a graph. It is a normal practice to plot the dependent variable on the Oy-axis and the independent variable on the Ox-axis. The points plotted in the space form a scatter diagram which is given below.

Now we have to draw a line called the line of regression which passes through these points, bearing equal number of points on both sides. Here also, it is

convenient to draw a straight line. Further, the points on each side should be at such a distance that the sum of their squares is the least. This is what is called the method of least squares.



(ii) REGRESSION LINES

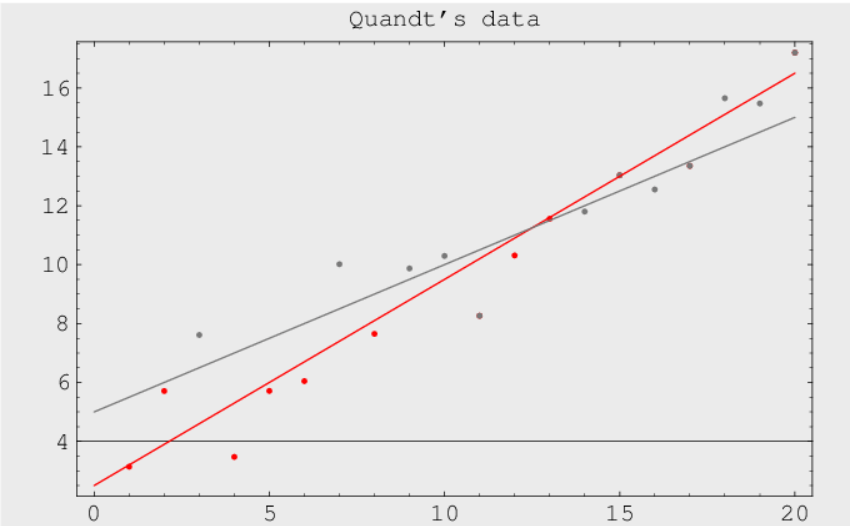
There are two regression lines ie., regression equation of Y on X and X on Y

The two linear equations are:

Y = a + bX……(i)

X = a+ bY……(ii)

where a is the intercept of the line and b the slope. The parameter b could have been indicated as b_{yx} and b, respectively, in the above two cases. In (i), Y is dependent variable and in (ii), X is the dependent variable. The above equations are based on the method of least squares. In the case of (i)the square of the vertical distances indicated in diagram (a) is the least and in case of (ii) the square of the horizontal distances indicated in the diagram (b) is the least.



For arriving at the regression equations we have to find out the values of the constants a and b. Let us do it with the help of normal equations.

For the line Y= a+ bX the regression equation of Y on X, the normal equations are:

$\Sigma Y = Na + b\Sigma X$

$\Sigma XY = a\Sigma X + b\Sigma Y^2$.

For the line X = a + bY, the regression equation of X on Y, the normal equations are

$\Sigma X = Na + b\Sigma Y$

$\Sigma XY = a\Sigma X + b\Sigma X^2$

By solving the above normal equations we get a and b values , Finally by substituting a and b values in regression equations (i) and (ii) , we get the required regression lines.

We will illustrate the use of these equations in the following example:

Total output in units	Number of persons employed
1	1
3	2
5	3
6	4
5	5

Calculation of Y^2 and XY will be made in the following manner :

X	Y	Y^2	XY
1	1	1	1
3	2	4	6
5	3	9	15
6	4	16	24
5	5	25	25
20	15	55	71

Therefore

$\Sigma X = Na + b\Sigma X$ ……..(i)

$\Sigma XY = a\Sigma X + b\Sigma X^2$ ……(ii)

By substituting values

20 = 5a+15b ……(iii)

71 = 15a+55b……(iv)

Multiplying (iii) above by 3 , we have

60 = 15a+45b……..(v)

Subtracting (iv) from (v) and after changing the sides, we have

15a +54b = 60

15a +55b=71

-10b =-11

10b = 11

b = 11/10 = 1.1

Substituting the value of b in equation (iii), we have

5ax15x11/10 = 20

Or 5a +33/2 = 20

Or 5a +16.5= 20

Or 5a =20 – 16.5 = 3.5

a = 3.5/5 = 0.7

Therefore the regression equation is X = 0.7+ 1.1Y

Similarly, we can find the constants of the regression Y = a+ bX

We can also arrive at standard expressions for the values of b and a by the use of the normal expressions as follows:

Normal equations:

$\Sigma Y = Na + b\Sigma X$ ……..(i)

$\Sigma XY = a\Sigma X + b\Sigma X^2$ ……..(ii)

$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$ ……..(iii)

$= \frac{\Sigma Y(\Sigma X)^2 - (\Sigma X Y)(\Sigma X)}{N\Sigma X^2 - (\Sigma X)^2}$……..(iv)

You will notice that the denominators of both the standard expressions are the same.

The numerator of (iii) is arrived at by taking the variables after remaining the normal equations as follows:

$Na + b\Sigma X = \Sigma Y$……..(v)

$a\Sigma X + b\Sigma X^2 = \Sigma XY$ ……..(vi)

The numerator of (iv) is arrived at by taking the cross multiplication of the variables in the normal equations after rearrangement as follows:

$Na = \Sigma Y - b\Sigma X$ ……..(v)

$a\Sigma X = \Sigma XY - b\Sigma X^2$ ……..(vi)

There exists an alternative way of finding the regression equations. Instead of using normal equations, deviations from the respective means or from assumed means are considered. An alternative way of using the original values also does exist. As a result, in many cases, computations become easier.

The regression equation of Y on X is given by

$Y - \overline{Y} = b_{xy} (X - \overline{X})$

And the regression equation of X on Y is given by

$X-\overline{X} = b_{xy}\ (Y-\overline{Y})$

Where$\overline{X}$, Y stand for respective means, bxy is the regression coefficient of X on Y, and b_{xy} is theregression coefficient of Y on X.

DEVIATIONS ($x = X-\overline{X}$, $y = Y-\overline{Y}$) TAKEN FROM REPSECTIVE MEANS ($\overline{X}$, $\overline{Y}$)

$b_{xy}= r\ \frac{\sigma_x}{\sigma_y} = \frac{\Sigma XY}{\Sigma Y^2}$ and $b_{yx}= r\ \frac{\sigma_y}{\sigma_x} = \frac{\Sigma XY}{\Sigma X^2}$

deviations $d_x = X- A$, and $d_y = Y- A$ taken from assumed means A

When actual means are not integers, for reasons of simplicity, we make use of arbitrary means and consider deviations d_x (the deviation in the variable X from the assumed means of X), d_y (the deviation in the variable Y from the assumed means of Y).

In this we use

$b_{xy} = r\ \frac{\sigma_x}{\sigma_y} = \frac{\Sigma dx\ dy-(\Sigma dx\ dy/N)}{\Sigma Y^2-[(\Sigma\ dy^2/N)]}$and $b_{yx}= r\ \frac{\sigma_y}{\sigma_x} = \frac{\Sigma\ dx\ dy-(\Sigma\ dx\ dy/N)}{\Sigma X^2-[(\Sigma dx^2/N)]}$

It is to be noted that the numerator of the expression $r\ \frac{\sigma_x}{\sigma_y}$ and that of $r\ \frac{\sigma_y}{\sigma_x}$is the same, butthe denominator is different.

USING ORIGINAL VALUES OF THE VARIABLES (without considering deviations)

In this case we use

$b_{xy}= r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma XY-(\Sigma XY/N)}{\Sigma Y^2-[(\Sigma Y^2/M)]}$ and $b_{yx}= r\ \frac{\sigma_y}{\sigma_x} = \frac{\Sigma XY-(\Sigma XY/N)}{\Sigma X^2-[(\Sigma X^2/N)]}$

Example : using the following data obtain the two regression equations

X	14	19	24	21	26	22	15	20	19
Y	321	36	48	37	50	45	33	41	39

Solution :

X	Y	$X=X-\overline{X}$	$Y=Y-\overline{Y}$	xy	X^2	Y^2
14	31	-6	-9	54	36	81
19	36	-1	-4	4	1	16
24	48	4	8	32	16	64
21	37	1	-3	-3	1	9
22	45	2	5	10	4	25
15	33	-5	-7	35	25	49
20	41	0	1	0	0	1
19	39	-1	-1	1	1	1
180	360	0	0	193	120	346

$\overline{X} = 180/9 = 20$ and $\overline{Y} = 360/9 = 40$

$b_{xy} = \Sigma XY / \Sigma Y^2 = 193/346 = 0.557$

and

$b_{yx} = \Sigma XY / \Sigma X^2 = 193/120 = 1.608$

regression equation of X on Y is given by

$X - \overline{X} = b_{xy} (Y - \overline{Y})$

Or $X - 20 = 0.557 (Y-40)$

i.e.$X = 0.557Y - 2.28$

regression equation of Y on X is given by

$Y - \overline{X} = b_{yx}(X - \overline{X})$

Or ,$Y - 40 = 1.608 (X-20)$

i.e.$Y = 1.608x + 7.84$

Example : from the following data find the regression equations and estimate the likely value of Y when X=100

X	72	98	76	81	56	76	92	88	49
Y	124	131	117	132	96	120	136	97	85

Solution :

X	Y	d_x=x-81	d_y=y-120	d_xd_y	d_x^2	dy^2
72	124	-9	4	-36	81	16
98	131	17	11	187	289	121
76	117	-5	-3	15	25	9
81	132	0	12	0	0	144
56	96	-25	-24	600	625	576
76	120	-5	0	0	25	0
92	136	11	16	176	121	256
88	97	7	-23	-161	49	529
49	85	-32	-35	1120	1024	1225
688	1038	-41	-42	1901	2239	2876

$\overline{X}$ = 688/9 = 76.44 and $\overline{Y}$ = 1038/9 = 115.33

$b_{xy}= \Sigma d_x d_y - (\Sigma d_x d_y/N)/ \Sigma y^2 - [(\Sigma d_y)^2/N]$

= 1901 –(-41x-42/9) / 2876 – $[(-42)^2/9]$

= 17109-1722/25884 – 1764

= 15387 / 24120

= 0.64

$b_{yx}= \Sigma d_x d_y - (\Sigma d_x d_y/N)/ \Sigma x^2 - [(\Sigma d_x)^2/N]$

= 1901 –(-41x-42/9) / 2239 – $[(-41)^2/9]$

= 17109-1722 / 20151 – 1681

= 15387 / 18470

= 0.83

regression equation of X on Y is given by

$X - \overline{X} = b_{xy} (Y - \overline{Y})$

Or X – 76.44 = 0.64 (Y-115.33)

i.e. X = 0.64Y + 2.63

regression equation of Y on X is given by

$Y - \overline{X} = b_{yx}(X - \overline{X})$

Or , Y – 115.33 = 0.83 (X-76.44)

i.e. Y = 0.83x + 51.88

Example: find the regression equations from the following data :

X= 60 Y=40 XY=1150

X^2 = 4160 Y^2 =1720 N=10

Solution:

$b_{xy} = \Sigma XY - (\Sigma X \Sigma Y/N) / \Sigma Y^2 - [(\Sigma Y)^2/N]$

= 1150-(60x40/10)/ 1720-[40^2/10]

=1150-240 / 1720 -160

= 910 / 1560

= 0.58

$b_{yx} = \Sigma XY - (\Sigma X \Sigma Y/N) / \Sigma X^2 - [(\Sigma X)^2/N]$

= 910 / 4160-360

= 0.24

regression equation of X on Y is given by

$X - \overline{X} = b_{xy} (Y - \overline{Y})$

Or X – (60/10)= 0.58[Y-(40/10)]

i.e. X = 0.58Y + 3.68

regression equation of Y on X is given by

$Y - \overline{X} = b_{yx}(X - \overline{X})$

Or , $Y - 4 = 0.24\ (X-6)$

i.e. $Y = 0.24X + 2.56$

PROPERTIES OF LINEAR REGRESSION

1. There are two regression equations as stated above, and the regression lines always intersect at the means ($\overline{X},\overline{Y}$).
2. Since b_{xy}, $b_{yx} = r^2$, r= b_{xy}, b_{yx}, and r, b_{xy}, b_{yx} , all have the same sign. The regression coefficients are zero if r =0.
3. Although regression equations are usually different, they become identical if r+1. It also follows from the regression equations that X=$\overline{X}$) and Y=$\overline{Y}$ if r=0, and these lines are perpendicular to each other.

A check on computational accuracy:

The important features of least squares regression line are (i) it passes through the point ($\overline{X},\overline{Y}$) corresponding to the mean of the observations X and Y , (i) the sum of the deviations of the Y's or the X's from their regression lines are zero i.e., where Y_c and X_c stand for the estimated values of the variables Y and X respectively.

$$(Y-Y)=0 \text{ and } \Sigma(X-X)=0$$

Thus, the positive and negative deviations about the regression line cancel so that the least square in goes through the centre of data-scatter-see the diagram below

The measure of variability based on the line of regression is derived from this basic concept, embodied in the method of least squares.