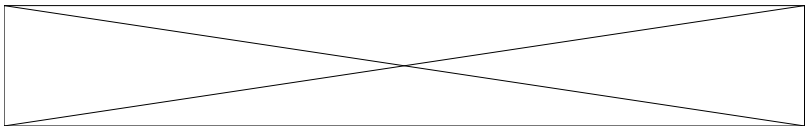
LESSON - 1

DEFINITION OF MATRIX



Please use headphones

Numbers arranged in rows and columns of the rectangular array and enclosed by square brackets [ ] or parenthesis ( ) or pair of double vertical line ||, || is called a matrix. Matrix is thus, defined as a rectangular array of ordered numbers in rows and columns. The numbers are also known as elements of matrix. A matrix consisting of m rows and n columns is written in the form as:

The first subscript 'm' refers to the 'number of rows' and the second subscript 'n' refers to the 'number of columns' in the particular matrix. Thus a_{12} is the element of 1^{st} row and 2^{nd} column in the matrix. Similarly, a_{21} is the element of 2^{nd} row and 1^{st} column, and aij refers to the element of i^{th} row and j^{th} column in the matrix. Generally, the dimension or order of a matrix is determined by the number of rows and columns it has. A matrix is denoted by capital letters like A, B, C etc. Some examples are:

$$A = \begin{bmatrix} 1 & -1 \\ 3 & 2 \end{bmatrix} \quad ; \quad 2x2 \text{matrix}$$

$$B = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 1 \end{bmatrix} \quad ; \quad 2x3 \text{matrix}$$

$$C = \begin{bmatrix} 11 & 2 & 10 & 1 \\ 2 & 4 & 10 & 1 \\ 10 & 3 & 5 & 6 \end{bmatrix} \quad ; \quad 3x4 \text{matrix}$$

2 x 2 matrix has 2 rows and 2 columns

2 x 3 matrix has 2 rows and 3 columns

3 x 4 matrix has 3 rows and 4 columns

Order of Matrix

If the matrix has mxn number of elements arranged in m rows and n columns, it is said to be of the order "m by n", which is written as m x n matrix. Remember that in a matrix...

a. Each element has its assigned position in row and column.

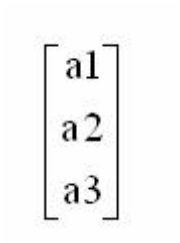
b. Two matrices of the same order are equal if the corresponding elements of them are same.

For example,

A = B if and only if $a_{11}=b_{11}$, $a_{12}=b_{12}$, $a_{21}=b_{21}$, and $a_{22}=b_{22}$.

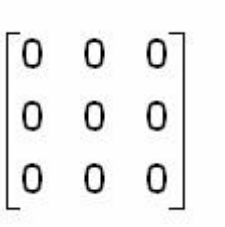
c. If the matrix consists of only one row as [a1 b1 c1], it is a Row Matrix or Row Vector.

d. If the matrix consists of only one column as



it is a Column Matrix or Column Vector.

e. A matrix is said to be a Zero Matrix or Null Matrix if and only if each of its elements is zero. For example,



f. A matrix containing the number of rows equal to the number of columns is known as a Square Matrix. For example,

is a square matrix of order 3 x 3, or a three-rowed square matrix.

In this, the elements 1, 9 and 8 are called the diagonal elements, and the diagonal is called the principal diagonal.

g. A matrix containing all diagonal elements as 'non-zero' and all other elements are zero is called diagonal matrix. For example,

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

h. In a diagonal matrix, if all the diagonal elements are equal, it is known as scalar matrix.
For example,

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

ALGEBRA OF MATRIX

(a) Additive of Matrix:

The addition of two or more matrices is possible only if they have the same order. The sum of matrices is obtained by adding the corresponding elements of the matrices.

Illustration 1

$$A = \begin{bmatrix} 3 & 1 \\ 4 & 2 \end{bmatrix} \text{ and } B = \begin{bmatrix} -2 & 5 \\ 6 & 0 \end{bmatrix}$$

Find A+B.

Solution

A + B =

If two matrices have different orders, their addition is not defined. For example, the following matrices cannot be added:

Properties of Matrix Addition :

i. Matrix addition is commutative, i.e. A + B = B + A

ii. Matrix addition is associative, i.e. (A + B) + C = A + (B + C)

iii. If for any matrix of A of dimension or order m x n, if there exists another matrix B of the same dimension, such that A + B = B + A = 0 (null matrix), then B is known as Additive Inverse or Negative of A, and is denoted by -A.

Illustration 2

Illustration 3

Explain the meaning of

Solution

The two matrices are of the same order 2x2, and are equal. This means that each element of one matrix is equal to the corresponding element of the second matrix. Therefore,

X+Y = 2

X+Z = 3 The relationship is a system of simultaneous equations.

Z-Y = 0

X+W = 4

Illustration 4

Reddy Company produces A, B and C products. The turnover (in 100 units) of these products for 1993 and 1994 regions are given below. Find total turnover for two years.

For 1993

Product	Region			
	N	E	W	S
A	29	36	18	13
B	24	16	28	26
C	10	13	18	21

For 1994

Product	Region			
	N	E	W	S
A	32	18	12	16
B	12	24	26	11

C	9	6	8	5

Solution

By additing both year's figures, the total turnover of each product for two years can be calculated:

For 1993 and 1994

Product	Region			
	N	E	W	S
A	29+32=61	36+18=54	18+12=30	13+16 =29
B	24+12=36	16+24=40	28+26=54	26+11 =37
C	10+9=19	13+6=19	18+8=26	21+5= 26

= 61 + 36 + 19 + 54 + 40 + 19 + 30 + 54 + 26 + 29 + 37 + 26 = 431 (100 units)

Total turnover is 43100 units.

Illustration 5

If $A = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 3 & -4 \\ 2 & 1 & 3 \end{bmatrix}$; $B = \begin{bmatrix} 3 & 1 & 4 \\ 1 & -2 & 2 \\ 1 & 0 & 1 \end{bmatrix}$ and $C = \begin{bmatrix} 1 & -2 & 2 \\ 2 & 1 & 2 \\ 0 & -1 & 2 \end{bmatrix}$

Find

(i) A + B + C;

(ii) (A + B) + C;

(iii) A + (B + C)

Solution

=

(ii) (A +B) +C =

=

(iii) A + (B + C) =

=

Illustration 6

Find A + B.

Solution

Illustration 7

Find A + B

solution

$$A+B=\begin{bmatrix}3 & -2\\ -1 & 4\end{bmatrix}+\begin{bmatrix}-3 & 2\\ 1 & -4\end{bmatrix}=\begin{bmatrix}3-3 & -2+2\\ -1+1 & 4-4\end{bmatrix}$$

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

Here, A + B = 0 (or B+ A = 0).

If A + B = 0, then B must be equal to -A, because A + (-A) = 0.

If A + B = 0, B is inverse or negative of A. That is, B = -A.

(B) Subtraction of Matrix

Like addition, the subtraction of two or more matrices is possible only if they have the same order. It is obtained by subtracting the corresponding elements of the given matrices.

Illustration 8

$$A=\begin{bmatrix}1 & 4\\ 2 & 1\end{bmatrix} \text{ and } B=\begin{bmatrix}2 & 3\\ 1 & 5\end{bmatrix}$$

Find (i) A - B, (ii) B - A, (iii) A + B, (iv) B + A. Comment.

Solution

$$(i)\ A-B=\begin{bmatrix}1 & 4\\ 2 & 1\end{bmatrix}-\begin{bmatrix}2 & 3\\ 1 & 5\end{bmatrix}=\begin{bmatrix}1-2 & 4-3\\ 2-1 & 1-5\end{bmatrix}=\begin{bmatrix}-1 & 1\\ 1 & -4\end{bmatrix}$$

$$(ii)\ B-A=\begin{bmatrix}2 & 3\\ 1 & 5\end{bmatrix}-\begin{bmatrix}1 & 4\\ 2 & 1\end{bmatrix}=\begin{bmatrix}2-1 & 3-4\\ 1-2 & 5-1\end{bmatrix}=\begin{bmatrix}1 & -1\\ -1 & 4\end{bmatrix}$$

$$(iii)\ A+B=\begin{bmatrix}1 & 4\\ 2 & 1\end{bmatrix}+\begin{bmatrix}2 & 3\\ 1 & 5\end{bmatrix}=\begin{bmatrix}1+2 & 4+3\\ 2+1 & 1+5\end{bmatrix}=\begin{bmatrix}3 & 7\\ 3 & 6\end{bmatrix}$$

$$(iv)\ B+A=\begin{bmatrix}2 & 3\\ 1 & 5\end{bmatrix}+\begin{bmatrix}1 & 4\\ 2 & 1\end{bmatrix}=\begin{bmatrix}2+1 & 3+4\\ 1+2 & 5+1\end{bmatrix}=\begin{bmatrix}3 & 7\\ 3 & 6\end{bmatrix}$$

Therefore A + B = B + A but A -B = B –A

(C) Multiplication of Matrix

The product of AB is got by multiplying the corresponding elements of row of A by those of column of B. In other words, the elements of row of AB is obtained by multiplying elements of row of A by the corresponding column elements of B and adding. For this, the number of elements in each row of A is equal to the number of elements in column B.

Multiplication of a matrix by a column matrix:

Illustration 9

Three persons intend to buy clothes at a shop, the prices of the clothes are given below:

Item	Price per unit
Pant	Rs. 80
Shirt	Rs. 60
Bush-shirt	Rs. 30
Tie	Rs. 20

Let us write the prices of these clothes in the form of a column, so that it becomes a column matrix:

Suppose person A wishes to buy 2, 2, 1, and 1 units of pant, shirt, bush-shirt, and tie respectively, this can be written as a row matrix as [2 2 1 1]. Then, the total amount of purchases by A can be obtained by multiplying the two matrices - matrix of units of clothes with the matrix of prices, as this...

= (2 x 80) + (2 x 60) + (1 x 30) + (1 x 20)

= 160 + 120 + 30 + 20

= Rs. 330.

If person B wishes to buy 1, 2, 2, and 1 units of pant, shirt, bush-shirt, and tie respectively, the total amount spent by B can be calculated by multiplying the following matrices...

$$B = \begin{bmatrix} 1 & 2 & 2 & 1 \end{bmatrix} \times \begin{bmatrix} 80 \\ 60 \\ 30 \\ 20 \end{bmatrix}$$

= (1 x 80) + (2 x 60) + (2 x 30) + (1 x 20)

= 80 + 120 + 60 + 20

= Rs. 280.

Similarly, if person C wishes to buy 3, 2, 2, 1 units of the respective clothe items, the total amount paid by C is...

$$C = \begin{bmatrix} 3 & 2 & 2 & 1 \end{bmatrix} \times \begin{bmatrix} 80 \\ 60 \\ 30 \\ 20 \end{bmatrix}$$

= (3 x 80) + (2 x 60) + (2 x 30) + (1 x 20)

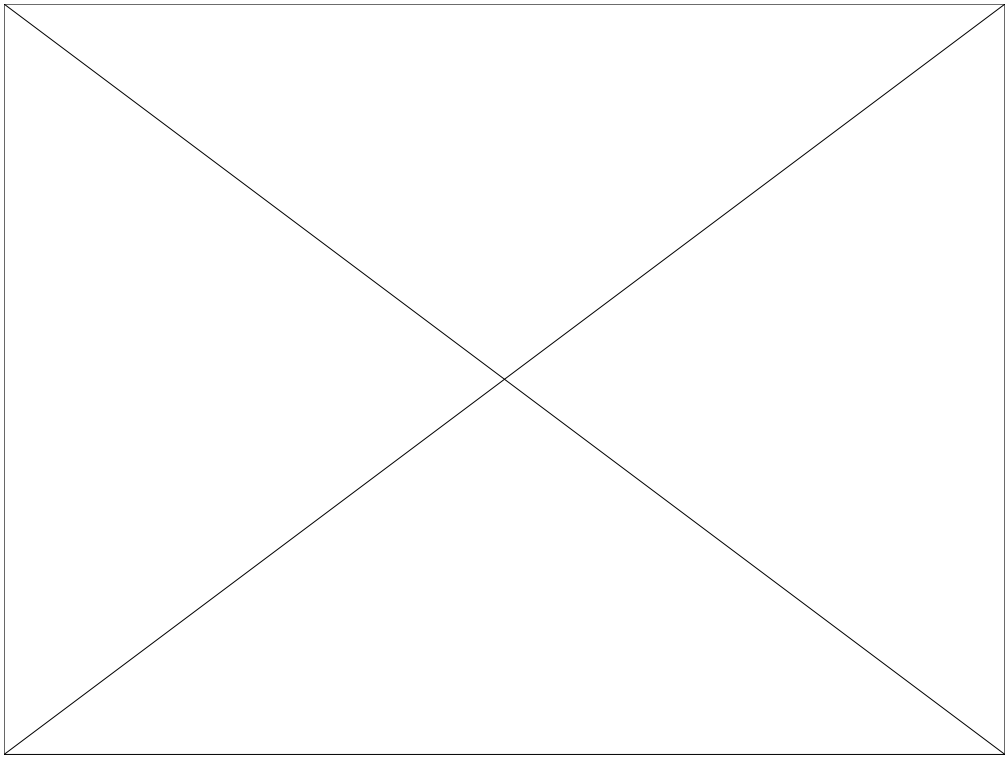
= 240 + 120 + 60 + 20

= Rs. 440.

Alternately, instead of doing three claculations separately, it can be done at one go as...

OR

=

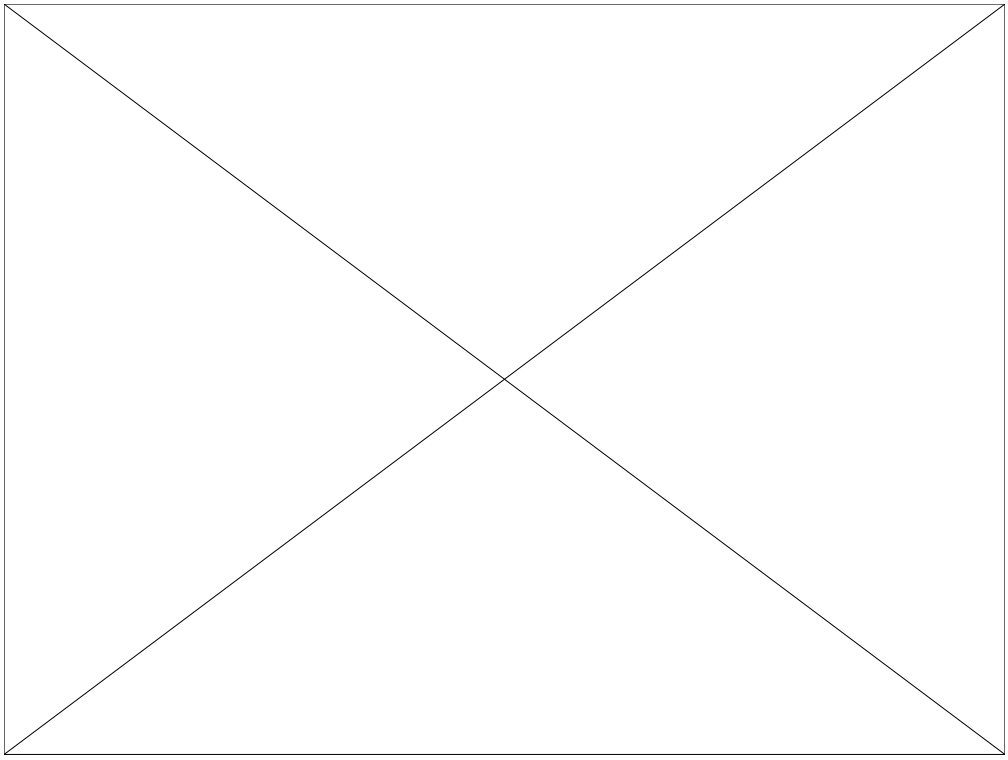


Please use headphones

- End of Chapter -

LESSON - 2

MULTIPLICATION OF A MATRIX BY ANOTHER MATRIX



Please use headphones

Illustration 10

Three persons intend to buy clothes at shops R_1, R_2, R_3 and R_4, the prices are given below.

Price Matrix:

Shops

Price	R_1	R_2	R_3	R_4
Pant (P)	50	70	60	40
Shirt (S)	40	60	30	40
Bush-shirt (B)	20	30	30	20
Tie (T)	30	20	20	30

Clothes Matrix:

	P	S	B	T
A	2	2	2	1

B	1	2	1	2
C	2	1	4	2

Assuming that a person should buy his lot from the same shop, which shop should A prefer.

Solution

Shops

$$\begin{matrix} R_1 & R_2 & R_3 & R_4 \end{matrix}$$

$$\begin{bmatrix} 50 & 70 & 60 & 40 \\ 40 & 60 & 30 & 40 \\ 20 & 30 & 30 & 20 \\ 30 & 20 & 20 & 30 \end{bmatrix}$$

$$\begin{matrix} P & S & B & T \end{matrix}$$

$$= \begin{bmatrix} 2 & 2 & 2 & 1 \\ 1 & 2 & 1 & 2 \\ 2 & 1 & 1 & 2 \end{bmatrix}$$

$$C = \begin{bmatrix} (2\times 50)+ & (1\times 40)+ & (1\times 20)+ & (2\times 30) \\ (2\times 70)+ & (1\times 60)+ & (1\times 30)+ & (2\times 20) \\ (2\times 60)+ & (1\times 30)+ & (1\times 30)+ & (2\times 20) \\ (2\times 40)+ & (1\times 40)+ & (1\times 20)+ & (2\times 30) \end{bmatrix} = \begin{bmatrix} 220 \\ 270 \\ 220 \\ 200 \end{bmatrix}$$

	R_1	R_2	R_3	R_4
A	**250**	**340**	**260**	**230**
B	**210**	**260**	**190**	**200**
C	**220**	**270**	**220**	**200**

First row gives the different amounts to be paid by A for the same lot at four (R1, R2, R3 , R4) shops; and so also second row and third row relate the amounts to be paid by B and C at four shops respectively.

For A and C, it would be better to buy the whole lot from shop R4, and B from shop R3 as the payment would be less.

Properties of Matrix Multiplication

(i) All the matrices cannot be multiplied with each other. Matrix multiplication in general is not commutative

i.e., A x B is not equal to B x A

Remark:

Given matrix A of order m x n, and matrix B of order p x q, the ordered pair of matrices A and B (A, B) is said to be conformable only if n = p (i.e. the number of columns of A = number of rows of B). If p and n are not equal, then AxB is not conformable. If AxB is conformable, it is not necessary that BxA must also

$$A = \begin{bmatrix} 2 & 4 & 6 \\ 3 & 5 & 8 \\ 4 & 2 & 3 \end{bmatrix} \text{ and } B = \begin{bmatrix} 8 & 2 \\ 4 & 5 \\ 2 & 3 \end{bmatrix}$$

conformable.

A is a 3x3 matrix, and B is a 3x2 matrix. So, m=3, n=3, p=3, q=2.

Ordered pair (A, B) is conformable because n = p = 3. Therefore AB is defined.

Is BA defined? No, because the number of columns of B (q = 2) is not equal to the number of rows in A (m = 3).

ii. Matrix multiplication is associative,

i.e., $A \times (B \times C) = (A \times B) \times C$

(when the order of matrices A, B, C are mxn, nxp, pxq respectively)

iii. Matrix multiplication is distributive

i.e., $A \times (B + C) = A \times B + B \times C$

(where the order of matrices A, B, C are mxn, nxp, pxq respectively)

Illustration 11

In family R, men, women and children are 3, 2, 1 respectively; and in family Q they are 1, 1, 2 respectively. The recommended daily calories are 2000, 1800 and 1500 for men, women and children, and proteins recommended are 50, 45, 35 respectively. Calculate the total requirement of calories as well as proteins for each of the two families.

Solution

Arrange the given information in matrix form.

Man Woman Child

Calorie Protein

The dimensions / order of two matrices A and B satisfy the rule of multiplication (no. of columns in A = no. of rows in B).

By multiplying A and B, we can get the total requirement of calories and proteins for R and Q families.

M W C Cal Prot

Illustration 12

A company produces three products A, B and C which it sells in two markets X and Y. Data given is...

$$\begin{array}{c} \\ X \\ Y \end{array} = \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 8000 & 3000 & 9000 \\ 5000 & 10000 & 7000 \end{bmatrix} \end{array}$$

If the unit sales price of A, B and C are Rs.3, Rs.2, and Rs.1 respectively, find the total revenue in X and Y market with matrix algebra.

Solution

Arrange the given information in matrix form.

The total revenue received by selling the products in markets X and Y is Rs.39000 and Rs.42000 respectively. The revenue of the company from both the markets is Rs.81000:

Illustration 13

If the unit cost of the above three products (given in Illustration 12) are Rs. 2.50, Rs. 1.75 and Rs. 0.80 respectively, find the profits.

Solution

Arrange the given data in matrix form...

Profit by selling products in X market and Y market is Rs. 6550 and Rs. 6400 respectively. Total profit earned by the PQR Company is Rs. 12950.

Illustration 14

A factory employs 40 skilled workers and 20 unskilled workers. The daily wages to skilled and unskilled are Rs. 35 and Rs. 22 respectively. Using matrices, find

(i) the matrix for number of workers

(ii) the daily payment made to them

Solution

Illustration 15

Find AX.

If AX = B, find X_1 and X_2

Solution

$$AX = \begin{bmatrix} 3X_1 + 2X_2 \\ 10X_1 - 2X_2 \end{bmatrix}$$

$$\text{If } AX = B, \text{we can write} \quad = \begin{bmatrix} 3X_1 + 2X_2 \\ 10X_1 - 2X_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 3 \end{bmatrix}$$

which means

$3X_1 + 2X_2 = 7$ (1)

$10X_1 – 2X_2 = 3$ (2)

Add (1) and (2)

$13X_1 = 10$

Put $X_1 = 0.77$ in (1)

$3(0.77) + 2X_2 = 7$

$2X_2 = 7 - 2.31$

$2X_2 = 4.69$

$X_2 = 2.345$

Illustration 16

$$If\ A = \begin{bmatrix} -3 & 2 \\ 5 & 0 \end{bmatrix} \quad B = \begin{bmatrix} -K & K \\ 2K & 1 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \quad D = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Find (i) AB, (ii) BC, (iii) CD, (iv) BD, (v) 2A + C, (vi) $(AD)^2$

Solution

$$(i)\ AB = \begin{bmatrix} (-3)(-K) + 2(2K) & (-3)K + 2(1) \\ 5(-K) + 0(2k) & 5(K) + 0(1) \end{bmatrix} = \begin{bmatrix} 7K & -3+2 \\ -5K & 5K \end{bmatrix}$$

$$(ii)\ BC = \begin{bmatrix} -K(0) + 5(0) & -K(1) + K(1) \\ 2K(0) + 1(0) & 2K(1) + 1(1) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 2K+1 \end{bmatrix}$$

$$(iii)\ CD = \begin{bmatrix} 0(1) + 1(0) & 6(0) + 1(0) \\ 0(1) + 1(0) & 0(0) + 1(0) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

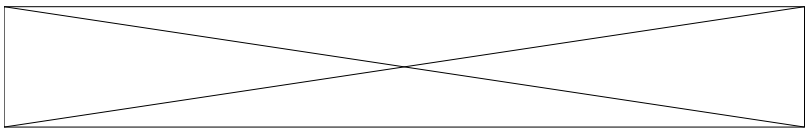
$$(iv)\ BD = \begin{bmatrix} -K(1) + K(0) & -K(0) + K(0) \\ 2K(1) + 1(0) & 2K(0) + 1(0) \end{bmatrix} = \begin{bmatrix} -K & 0 \\ 2K & 0 \end{bmatrix}$$

$$(v)\ 2A + C = \begin{bmatrix} -6+0 & 4+1 \\ 10+0 & 0+1 \end{bmatrix} = \begin{bmatrix} -6 & 5 \\ 10 & 1 \end{bmatrix}$$

$$(iv)\ (AD)^2 = AD \times AD$$

$$AD = \begin{bmatrix} -3(1) + 2(0) & -3(0) + 2(0) \\ 5(1) + 0(0) & 5(0) + 0(0) \end{bmatrix} = \begin{bmatrix} -3 & 0 \\ 5 & 0 \end{bmatrix}$$

$$AD^2 = \begin{bmatrix} -3(-3) + 0(5) & -3(0) + 0(0) \\ 5(-3) + 0(5) & 5(0) + 0(0) \end{bmatrix} = \begin{bmatrix} 9 & 0 \\ -15 & 0 \end{bmatrix}$$



Please use headphones

D. Symmetric and Skew Symmetric Matrices:

Square matrix A is symmetric matrix of $a_{ij} = a_{ji}$ for all the values of i & j

Example

$$\begin{bmatrix} 3 & 6 & 9 \\ 6 & 5 & 7 \\ 9 & 7 & 4 \end{bmatrix} \quad \begin{bmatrix} a & b & -c \\ b & e & -g \\ -c & -g & f \end{bmatrix}$$

A square matrix A is said to be skew symmetric if $a_{ij}=-a_{ij}$ for all the values of i & j

Example

$$\begin{bmatrix} o & g & h \\ -g & o & f \\ -h & -f & o \end{bmatrix}$$

Transpose of a matrix:

If A is a matrix, the matrix obtained by changing its rows into columns & columns into rows is known as transpose of the matrix, and denoted by A' or A^T

$$If \quad A = \begin{bmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \end{bmatrix}_{3\times 2}$$

$$then \quad A^T = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \end{bmatrix}_{2\times 3}$$

Illustration 17

Prove that (A+B)' = A' + B'

Let

$$A = \begin{bmatrix} 2 & 4 \\ 5 & 3 \end{bmatrix} \quad and \quad B = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix}$$

$$A+B = \begin{bmatrix} 3 & 7 \\ 9 & 5 \end{bmatrix} \quad and \quad (A+B)' = \begin{bmatrix} 3 & 9 \\ 7 & 5 \end{bmatrix}$$

$$A' = \begin{bmatrix} 2 & 5 \\ 4 & 3 \end{bmatrix} \quad and \quad B' = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}$$

$$A' + B' = \begin{bmatrix} 3 & 9 \\ 7 & 5 \end{bmatrix}$$

$$A'+B' = (A+B)'$$

Illustration 18

Prove that (AB)' = B'A'

$$Let\ A = \begin{bmatrix} 2 & 4 \\ 5 & 3 \end{bmatrix} \quad and \quad B = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix}$$

$$AB = \begin{bmatrix} 18 & 14 \\ 17 & 21 \end{bmatrix} \quad and (AB)' = \begin{bmatrix} 18 & 17 \\ 14 & 21 \end{bmatrix}$$

$$A' = \begin{bmatrix} 2 & 5 \\ 4 & 3 \end{bmatrix} \quad and \quad B' = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}$$

$$B'A' = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix} \times \begin{bmatrix} 2 & 5 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 18 & 17 \\ 14 & 21 \end{bmatrix}$$

Therefore

$$(AB)' = \begin{bmatrix} 18 & 17 \\ 14 & 21 \end{bmatrix} = B'A' = \begin{bmatrix} 18 & 17 \\ 14 & 21 \end{bmatrix}$$

- End of Chapter -

LESSON - 3

DETERMINANT OF A MATRIX

A single number expressing the difference between two or more products is called determinant (which is denoted by either | A | or det. A). It is represented by a group of numbers enclosed by two vertical lines. For example, a, b, c, d are the elements of the determinant...

$$\det \begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

and its value is (a x d) - (b x c)... the difference between the products of the elements of the two diagonals.

Determinants are possible only for square matrices.

Illustration 19

$$= 6 \times 4 - 8 \times 2$$

$$= 24 - 16 = 8$$

The determinant of a square matrix is of order 1, 2, 3, and so on. Let us define the two important methods which are used to find the determinants to matrices of order 3 & order 4.

Minor

The 'minor of an element' in a determinant is the determinant of one lower order obtained by deleting the row and the column containing that element. Thus, in the determinant of order 3.

The Minor (M) of the element a_{mn} can be obtained by deleting m^{th} row and n^{th} column. So, in the given matrix, minor of element a_{11} will be obtained by deleting 1^{st} row and 2^{nd} column of the matrix...

Cofactor

The Cofactor C_{ij} of an element a_{ij} is defined as $(-1)^{i+j} M_{ij}$, where M_{ij} is the minor of the element a_{ij}.

Illustration 20

Find the value of determinant of $\begin{vmatrix} 1 & 20 & 70 \\ 2 & 40 & 90 \\ 2 & 45 & 75 \end{vmatrix}$

Solution

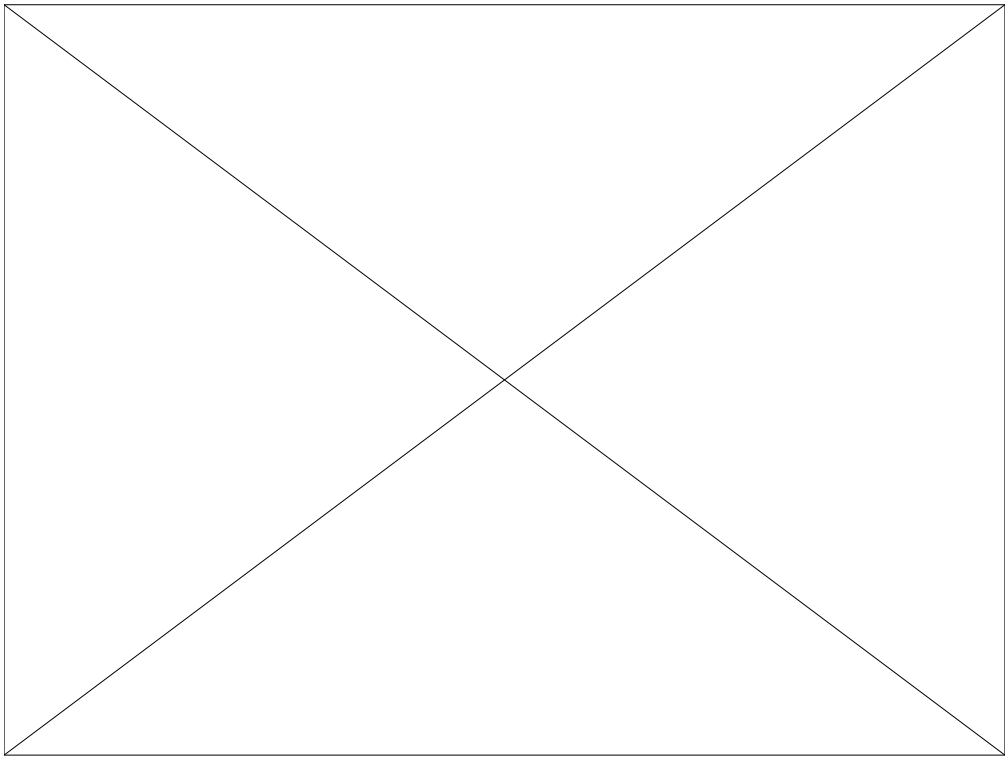
Expand the determinant by using the elements of the first column. We get:

= 1 (40 x 75 - 45 x 90) -2 (20 x 75 - 45 x 70) + 2 (20 x 90 - 40 x 70)

= 1 (3000 – 4050) -2 (1500 – 3150) + 2 (1800 – 2800)

= -1050 + 330 - 2000

= 250



Please use headphones

Cramer's Rule

In this rule, the determinant is known by two simultaneous equations. To solve the equations

(i) $a_1x + b_1y = m_1$ and

(ii) $a_2x + b_2y = m_2$,

multiply equation (i) with b_2 and equation (ii) with b_1, to get

$$a_1b_2x + b_1b_2y = m_1b_2 \quad ...1$$

$$a_2b_1x + b_2b_1y = m_2b_1 \quad ...2$$

Subtract equation 2 from equation 1 to get

$$(a_1b_2x - a_2b_1x) + (b_1b_2y - b_1b_2y) = m_1b_2 - m_2b1$$

$$x(a_1b_2 - a_2b_1) = m_1b_2 - m_2b_1$$

$$x = \frac{m_1b_2 - m_2b_1}{a_1b_2 - a_2b_1}$$

Similarly, multiply equation (i) with a_2 and equation (ii) with a_1, to get

$$a_1a_2x + b_1a_2y = m_1a_2 \quad ...3$$

$$a_2a_1x + b_2a_1y = m_2a_1 \quad ...4$$

and subtract equation 4 from equation 3 to get

$$(a_1a_2x - a_2a_1x) + (b_1a_2y - b_2a_1y) = m_1a_2 - m_2a_1$$

$$y(b_1a_2 - b_2a_1) = m_1a_2 - m_2a_1$$

$$y(b_2a_1 - b_1a_2) = m_2a_1 - m_1a_2$$

$$y = \frac{m_2a_1 - m_1a_2}{a_1b_2 - a_2b_1}$$

Solution for x and y in matrix form is...

No solution exists if the denominator is zero.

Illustration 21

Solve for x and y by using determinants

$4x + 2y = 2$

$3x – 5y = 21$

Solution

We have,

$a_1 = 4, \quad b_1 = 2, \quad m_1 = 2$

$a_2 = 3, \quad b_2 = -5, \quad m_2 = 21$

So,

Illustration 22

Find x, y and z for the following equations.

$x + 6y – z = 10$

$2x + 3y + 3z = 17$

$3x - 3y - 2z = - 9$

Solution:

For equations,

$a1x + b1y + c1z = m1$

$a2x + b2y + c2z = m2$

$a3x + b3y + c3z = m3$

solution is

Hence, the solution for the given equations is:

$$x = \frac{\begin{vmatrix} 10 & 6 & -1 \\ 17 & 3 & 3 \\ -9 & -3 & -2 \end{vmatrix}}{\begin{vmatrix} 1 & 6 & -1 \\ 2 & 3 & 3 \\ 3 & -3 & -2 \end{vmatrix}} = \frac{96}{96} = 1$$

$$y = \frac{\begin{vmatrix} 1 & 10 & -1 \\ 2 & 17 & 3 \\ 3 & -3 & -2 \end{vmatrix}}{\begin{vmatrix} 1 & 6 & -1 \\ 2 & 3 & 3 \\ 3 & -3 & -2 \end{vmatrix}} = \frac{192}{96} = 2$$

$$z = \frac{\begin{vmatrix} 1 & 6 & 10 \\ 2 & 3 & 17 \\ 3 & -3 & -9 \end{vmatrix}}{\begin{vmatrix} 1 & 6 & -2 \\ 2 & 3 & 3 \\ 3 & -3 & -2 \end{vmatrix}} = \frac{288}{96} = 3$$

The useful properties of determinants of any order are:

1. The value of a determinant remains unchanged even if columns are changed into rows, and rows changed into columns.
2. If two columns (or rows) of a determinant are interchanged, the value of the determinant so obtained is negative of the value of the original determinant.
3. If each element in any one row or any one column of a determinant is multiplied by a constant, say K, then the value of determinant so calculated is K times the value of the original determinant.
4. If any two columns (or any two rows) in a determinant are equal, the value of the determinant is zero. Such a determinant is said to be linearly dependant, otherwise it is called independent.

Inverse of Matrix

In matrix theory, the concept of dividing one matrix directly by another does not exist. However, a unit matrix can be divided by any square matrix by a process called "inversion of a matrix". In algebra if X x Y = 1, then X = 1/Y or we say that Y is the inverse of X, or X is inverse of Y. In matrices, inverse of a matrix A is represented by A^{-1}. Product of A and A^{-1} must be equal to identity matrix (denoted by I).

$A \times A^{-1} = I$

$A^{-1} = I / A$

The inverse matrix concept is very useful in solving simultaneous equations in input-output analysis as well as regression analysis in economics.

Illustration 23

Find the inverse of matrix

Solution

$= 3x2 - 4x1 = 6 - 4 = 2$

(Adj A is obtained by interchanging the principal diagonal elements, and revesing signs of the elements of the other diagonal in A)

$A^{-1} = \text{Adj } A / \text{Det } A,$

So,

Remember that:

1. The inverse of the inverse is the original matrix i.e. $(A^{-1})^{-1} = A$.
2. The inverse of transpose of a matrix is the transpose of its inverse, i.e. $(A^{-1})^T = (A^T)^{-1}$
3. The inverse of the product of two non-singular matrices is equal to the product of two inverses in the reverse order i.e. $(AB)^{-1} = B^{-1} A^{-1}$

Rank of Matrix

The rank of a matrix is the maximum number of linearly independent (non-singular) rows (or columns) in the matrix. The rank of a matrix A is the order of the largest non-zero minor of |A|.

The following points should be kept in mind:

1. The rank of matrix is not related to any way to the number of zero elements in it.
2. The rank cannot exceed the number of its rows or columns, whichever is lesser.
3. The rank is at least 1, unless the matrix has all zero elements.
4. The rank of a column matrix (single column) having any number of rows, say m x 1 matrix, is at most 1; rank of a 3 x 50 matrix is at most 3.
5. Rank of the transpose of matrix A is the same as that of A.

Note:

The straight way to find the rank of any matrix is to look for non-zero determinant of the highest order which the given matrix contains.

Illustration 24

Find the rank of

Solution

We see in the above calculation that each minor is non-zero. Therefore, the rank can be at most 2.

We find that there is atleast one non-zero determinant of order 2:

Hence the rank of matrix A is 2.

Now,

Further, each of the 2nd order determinant is also zero.

Again, the matrix is not null matrix (it's elements are non-zero). So, the 1st order determinants are non-zero.

Hence the rank of matrix B is 1.

Simultaneous Equations

Matrix algebra is useful in solving a set of linear simultaneous equations involving more than two variables. The procedure for getting the solutions is as :

Consider the set of linear simultaneous equations:

$$x + y + z = 4$$

$$2x + 5y - 2z = 3$$

Let us write the system of equations in matrix form.

or AX = B, where

A is known as the coefficient matrix in which coefficients of X are written in the first column, coefficients of y in the second column, and coefficients of z in the third column.

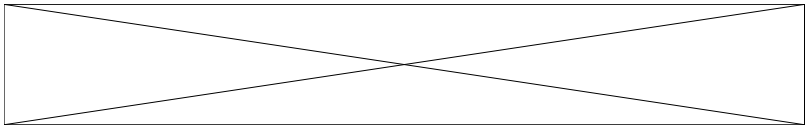
X is the matrix of unknown variable x, y and z.

B is the matrix formed with the right hand terms of the equations.

Note:

A is a 2x3 matrix, X is a 3x5 matrix, and B is a 2x1 matrix.

Linear Equations



Please use headphones

If matrix B is zero then the system AX = 0, which is said to be homogeneous system, otherwise the system is said to be non-homogeneous.

To solve the equation AX = 0, an elementary operation on the coefficient matrix A should be done. This elementary operation is called row operation if it applies to rows, and column operation if it applies to columns.

The elementary operation is of any one of the three types —

(1) Interchange of any two rows (or columns),

(2) Multiplication (or division) of the elements of any row (or column) by any non-zero number,

(3) Addition of the elements of any row (or column) to the corresponding elements of any other row (or column) or multiplied by any number.

To solve homogeneous linear equations, the Guass-Jordon method also called Triangular Term Reduction method is applied. In this, the given linear equation is reduced to an equivalent simpler system, which is studied in both homogeneous and non homogeneous equations. The simplified system is as

$$x_1 + b_1x_2 + c_1x_3 = d_1$$

$x_2 + c_2x_3 = d_2$

$x_3 = d_3$

The non-homogeneous linear equations can be solved by either (i) Matrix method, or (ii) Cramer's method or (iii) Gauss Jordon method.

Matrix Method

Let AX = B be the given system of linear equations, and A^{-1} be the inverse of A. Pre-multiplying both sides of the equation with A^{-1}, we get

$A^{-1} (AX) = A^{-1} B$

$(A^{-1}A) X = A^{-1} B$

$I X = A^{-1} B$

$X = A^{-1}B / I = A^{-1}B$

(I is the identity matrix, and its value is 1)

X gives the solution to the given set of simultaneous equations. It is thus, calculated by first finding A^{-1} and then post-multiplying A^{-1} with B.

Illustration 25

Solve the equations and find the values by linear equations using the matrix inverse method.

Solution

The equations for three days cost can be written like this:

Given data in matrix form, AX = B.

Since Det A is not zero, inverse of matrix A exists. It is computed as

A^{-1} = Adj A / Det A

Adj A is the matrix of Cofactors of elements of A

$X = A^{-1}B$

Proof

a+10b + 40c = 6950

5000+10´75 ´40´30 =6950

5000+750+1200=6950

6950=6950

Illustation 26

Find out the cars of each type to be produced.

Type	C1	Car C2	C3	Material Available
R1	2	3	4	29
R2	1	1	2	13
R3	3	2	1	16

Solution

Denote C1, C2 and C3 by X1, X2 and X3 respectively. The system of three linear equations become

$$\begin{bmatrix} 2X_1 + & 3X_2 + & 4X_3 \\ X_1 + & X_2 + & 2X_3 \\ 3X_1 + & 2X_2 + & X_3 \end{bmatrix} = \begin{bmatrix} 29 \\ 13 \\ 16 \end{bmatrix}$$

This in matrix form is as:

$$\begin{bmatrix} 2 & 3 & 4 \\ 1 & 1 & 2 \\ 3 & 2 & 1 \end{bmatrix} \times \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 29 \\ 13 \\ 16 \end{bmatrix}$$

Applying Cramer's Rule

$$X_1 = \frac{\begin{vmatrix} 29 & 3 & 4 \\ 13 & 1 & 2 \\ 16 & 2 & 1 \end{vmatrix}}{\begin{vmatrix} 2 & 3 & 4 \\ 1 & 1 & 2 \\ 3 & 2 & 1 \end{vmatrix}} = \frac{29(-3) - 3(-19) + 4(10)}{2(-3) - 3(-5) + 4 + (-1)} = 2$$

$$X_2 = \frac{\begin{vmatrix} 2 & 29 & 4 \\ 1 & 13 & 2 \\ 3 & 36 & 1 \end{vmatrix}}{\begin{vmatrix} 2 & 3 & 4 \\ 1 & 1 & 2 \\ 3 & 2 & 1 \end{vmatrix}} = \frac{2(-19) - 29(-5) + 4(-23)}{2(-3) - 3(-5) + 4(-1)} \frac{15}{5} = 3$$

$$X_3 = \frac{\begin{vmatrix} 2 & 3 & 29 \\ 1 & 1 & 13 \\ 3 & 2 & 36 \end{vmatrix}}{\begin{vmatrix} 2 & 3 & 4 \\ 1 & 1 & 2 \\ 3 & 2 & 1 \end{vmatrix}} = \frac{2(10)-3(-3)+29(-1)}{2(-3)-3(-5)+4(-1)} = 4$$

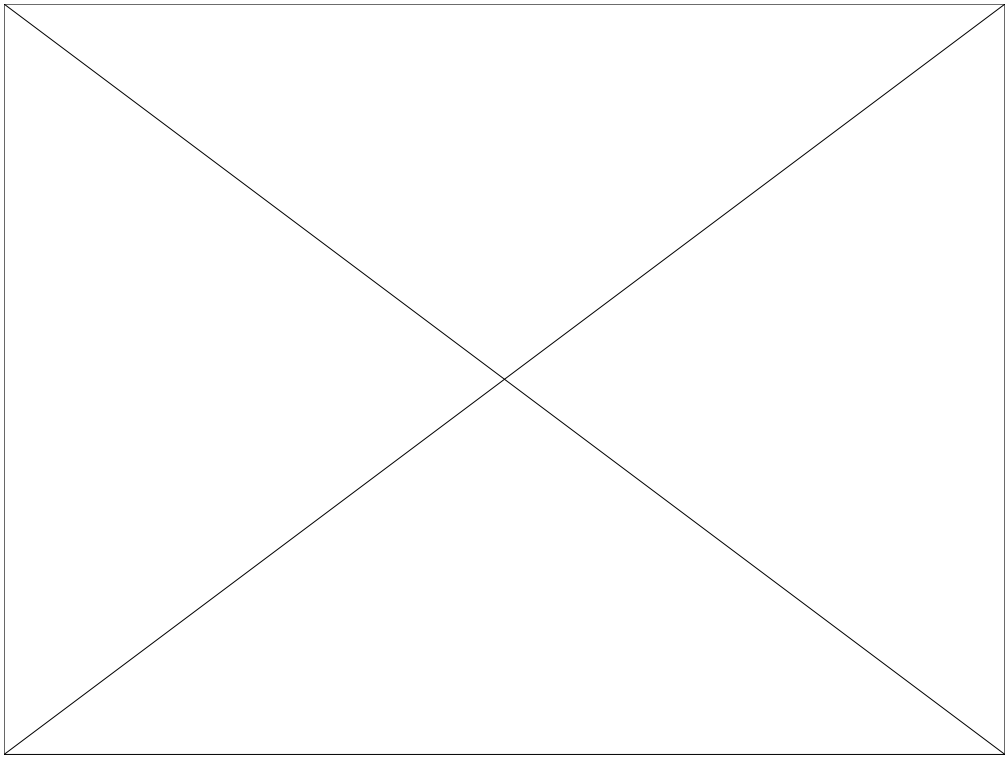
The number of cars of type C1, C2, and C3 to be produced are 2, 3 and 4 respectively.

- End of Chapter -

LESSON - 4

SET THEORY

A set is fundamental and all mathematical objects as well as constructions need the set theory. In many fields, scientists plan for construction of their objects in terms of sets. In trade, industry and commerce analysis, we have sets of data, sets of items produced, sets of outcomes of decisions and alike. While expressing the words such as family, association, group, crowd, we often used to convey the idea of a set in our day life. Enumerations and calculations leading to numbers form set and which gives a good insight into the depth of nature. In set, symbols are used, on which its operations can be done.



Please use headphones

Notation

A collection of any type of numbers, things, or objects is referred to "set". The constituent numbers of it is termed as its "elements". These elements may be presented by enclosing them in brackets or may be described in a form with statement of the properties. The sets is denoted by capital letters A, B, C,.....and and their elements in lower case letters a, b, c,..... A set is known by its elements. A set may be presented either in tabular method (or Roster method), descriptive phrase method or rule method (or set builder). These are discussed with examples.

A = {2, 4, 6, 8} is a set of the even numbers 2,4, 6, 8. It means A is a set of all even numbers between 1 and 9; and 2, 4, 6, 8 are elements of A. The function of this set is described in tabular form enlisting each element of it within the bracket. This method is known as tabular method.

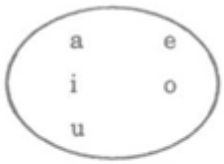
A = {even numbers between 1 and 9}. In this, the elements of A simply place a phrase describing the elements of the set within the bracket. This method is called descriptive phrase method.

A = {x/x is an even integer, $1 < x < 0$}. In this, the set is determined by its elements but not by the description. This method is known as rule method. However, it should be noted that the set A we get in any of the three different ways as shown above, remains the same.

Venn Diagram

A diagram is said to be Venn diagram if the elements of set are represented by points in circular or similarly enclosed curve. For example, V = {a, e, i, o, u}. The Venn diagram shows the letters a, e, i, o, u which are the elements of set V. A set remains unchanged even if the order of its elements is changed. Thus, the above set can be written as:

V = {e, o, i, u, a}



Venn diagram

Illustration 28:

Express the following in the set notation:

i. Integers less than 5

ii. The letters in the word 'English'

iii. Integers greater than 199 and less than 201.

iv. Even numbers up to 100

v. The three smallest integers greater than 20, and three smallest integers less than 20.

Solution

i. {0 < x < 5}

ii. {e, n, g, l, i, s, h}

iii. {200}

iv. {2, 4, 6, 8, 10, ..., 100}

v. {21, 22, 23} and {17, 18, 19}

Illustration 29

Explain the following sets

i. A = {x /x is a college having more than 200 students}

ii. B = {b /b>4}

iii. C = {c (integer)/c<0}

Solution

i. Set A consists of the elements x such that, x is a college having more than 200 students.

ii. Set B consists of the elements b such that b is all the numbers greater than 4.

iii. Set C consists of the elements c such that c is all integers less than 0 i.e. C is the set of all negative integers.

Finite and Infinite Sets

If A set has a finite number of elements, A is said to be a finite set. The number of elements is denoted by n(A) and can be counted by a finite number. A set is infinite, if it has an infinite number of elements. The elements of such set cannot be counted by a finite number. A set of points along a line or in a plane is known as point set.

Illustration 30

State the following in terms of finite or infinite sets

i. A= {a, b, c, y, 1, 4, 6, r}

ii. D = {x/x pages in a book}

iii. P = {x/x is a natural number}

iv. B = {x/x is a number between 5 and 6}

v. Is M = {x/x is a currency note in India} an infinite set?

vi. Is K = {x/x is a multiple of 3} an infinite set?

Solution

i. A is a finite set, and n(A) = 8

ii. D is a finite set

iii. P is an infinite set, {1, 2, 3, ..., n}

iv. B is a finite set

v. No

vi. Yes

Types of Sets

a. A set which contains no elements is called a null set or empty set, and is denoted by Ø (Phi)

b. A set which contains only one element is known as unit set.

c. A set which contains the totality of elements is called universal set. It is often drawn as a rectangle and is denote by E

d. If the two sets have no elements in common, they are disjoint sets.

e. If the two sets overlap, that overlapping portion will include the common points between the two sets is known as overlapping set.

f. If an element x belongs to a set A, then it is said to be the membership set.

g. If some elements in a set have a common special property, they can be said to belong to a subset.

h. If each and every element of one set is also an element of another set, the two sets are said to be equal or identical sets.

i. Two sets are said to be equivalent if there is a one-to-one correspondence between the elements in two sets. The elements appearance order in each set is immaterial. Equivalence may be expressed symbolically as, A = B or A⇔B

j. The class of all subsets of a set A is called the power set of A. It is denoted by P(A).

Illustration 31

State the following sets:

i.	A	=	{x /x is a married bachelor}
ii.	A	=	{x/x is a tomato growing a mango}
ii.	A	=	{a}
iv.	A	=	{0}
v.	E	=	{a, b, c, d, ..., z}
vi.	E	=	{tail, head}
vii.	E	=	{rise in supply, constant in supply, fall in supply }
viii.	A	=	{odd numbers} and B = {even numbers}
ix.	A	=	{0, l, 2, 3} and B = {5, 6}
x	A	=	{0, 1, 2, 3, 4, 5, 6, 8} and B ={2, 4, 6, 8}
xi	If	x	belongs to a set A and if x does belong to a set A
xii	A	=	{3, 5, 7} and B= {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
xiii	A	=	{commerce, economics} and B = {chemistry, mathematics, commerce, economics}
xiv	A	=	{2, 4, 6, 8} and B = {4, 8, 2, 6}
xv	A	=	{x/x is a letter of the word total and B {otalt}

Solution

i A = {x/x is a married bachelor } = Ø

ii A = {x/x is a tomato growing mango} = Ø

iii. A = {a} is a unit set, also called singleton set

iv. A = {0} is a unit set containing an element zero. A is not a null set

v. E = {a, b, c, d, ..., z} is a universal set (it covers all the letters of the English alphabet)

vi. E = {tail, head} is a universal set (it covers all the possibilities of tossing a fair coin)

vii. E = {rise in supply, constant in supply, fall in supply} is a universal set (it covers all the possibilities)

viii. If A = {odd numbers} and B = {even numbers}, then A and B are disjoint sets

ix. If A = {0, 1, 2, 3} and B = {5, 6}, then A and B are disjoint sets

x. A = {0, 1, 2, 3, 4, 6} and B {2, 4, 6, 8}, the sets are overlapping (the common points are 2, 4, 6)

xi. If x belongs to a set A, and if x does not belong to a set A, then we write x ¢ A, and x € A respectively

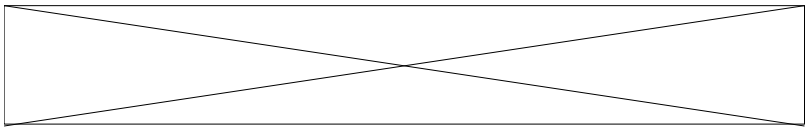
xii. A = {3, 5, 7} and B = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, then A is subset of B

xiii. A = {commerce, economics} and B = {chemistry, mathematics, commerce, economics}, then A is a subset of B

xiv. If A = {2, 4, 6, 8} and B = {4, 8, 2, 6}, then A = B

xv. If A = {x/x is a letter of the word total} and B = {otalt} then, A = B

Operations on Sets



Please use headphones

A set may be combined and operated in various ways to get new sets. The basic operations on sets are classified into four. They are (a) Complementation, (b)

Difference, (c) Intersection and (d) Union. These are discussed below with suitable examples.

a. Complementation

If A is a subset of the universal set E, then the set of elements of E that do not belong to A, is known as the complement of A.

The unshaded portion of of E in the above Venn diagram is referred to the complement of A. Therefore complement of a set A in a universal set E, is the set of all elements in E which are not elements of A. The complement of A may be written as A^1.

Illustration 32

Find the complement of A, given the universal set as E = {1, 2, 3, 4, 5, 6}, if:

i. A = {2, 4, 6}

ii. A = {1, 2, 5, 7}

Solution

i. If E = {1, 2, 3, 4, 5, 6} and A = {2, 4, 6}, then A^1 = {1, 3, 5}

ii. If E = {1, 2, 3, 4, 5, 6} and A = {1, 2, 5, 7}, then A^1 = {3, 4, 6}

b. Difference

Difference between the two sets A and B is the set of all elements belong to A but not B.

A - B = {x /x € A, x ¢ B} (x belongs to set A but does not belong to set B)

B - A = {x /x € B, x ¢ A} (x belongs to set B but does not belong to set A)

Illustration 33

If A = {a, b, x, y} and B = {c, d, x, y}, find A-B and B-A.

Solution

To get A - B, take set A = {a, b, x, y} and delete from it the elements that are present in B, which are x and y.

Hence, A-B = {a, b, ~~x~~, ~~y~~} = {a, b}

Similarly, to get B - A, take set B = {c, d, x, y} and delete from it the elements that are present in A, which are x and y.

Hence, B-A = {c, d, ~~x~~, ~~y~~} = {c, d}

c. Intersection

The intersection of two sets A and B is the set of all elements which belong to both A and B. Symbolically,

A(¯)B = {x / x € A and x € B}

Illustration 34

If A = (a, b, c, d} and B = (c, d, e}, find A(¯)B

Solution

The elements that are present in both A and B are {c, d}. Hence, A(¯)B = {c, d}

d.Union

The Union of two sets A and B is the set of all elements that belong to A or B or both A and B.

Symbolically, A U B = {x / x € A or x € B or x € both A and B}

Illustration 35

Given A = {0, 1, 2} and B = {2, 3, 4}, find

i) A(¯)B

ii) AUB

Solution

i) A(¯)B = {2} (elements that are present in both A and B)

ii) AUB = {0, 1, 2, 3, 4} (elements that are present in A and B put together)

Illustration 36

Find A(¯)B, A-B, and AUB, given

i) A = {c, f} and B = {c, d, g}

ii) A = {c, d} and B = {e, f, g}

Solution

i) If A = {c, f} and B = {c, d, g}

A(¯)B = {c}

A-B = {f}

AUB = {c, d, f, g}

ii) If A = {c, d} and B = {e, f, g}

A(¯)B = Ø

A-B = {c, d}

AUB = {c, d, e, f, g}

Exercise

1. Define matrix. Explain the operations on matrices.

2. What is determinant of a square matrix? Discuss its properties.

3. What is inverse of a matrix? Explain it with example.

4. Explain how the matrix algebra is useful in solving a set of linear simultaneous equations.

5. Find the rate of commission on items A, B, C from the data given below:

Month	A	B	C	Commission Rs.
April	90	100	20	900
May	50	50	30	800
June	40	70	20	650

6. If

prove that P=2.50, Q = 1.25 and R = 1.50.

7. What is a Venn diagram? Explain complement, union, difference and intersection with Venn diagrams.

8. Explain the following:

i. Addition of matrices

ii. Transpose of matrix

iii. Minor and Cofactor

iv. Cramer's rule

v. Vector matrix

vi. Rank of matrix

vii. Zero matrix

viii. Symmetric matrix

ix. Set

REFERENCES

Birkhoff, G. and Maclane, PA., 'Survey of Modern Algebra', The Macmillan Co.

Mehta and Madnani, 'Mathematics for Economies', Sulthan Chand and Sons, New Delhi, 1979.

Monga, G.S., 'Mathematics and Statistics for Economies', Vikas Publishing House, New Delhi, 1982.

Raghavachari, M. 'Mathematics for Management - An Introduction', Tata McGraw-Hill, Delhi, 1985.

Scarles, S.R. 'Matrix Algebra useful for Statistics', John Wiley and Sons, New York, 1982.

- End of Chapter -

LESSON - 5

PROBABILITY

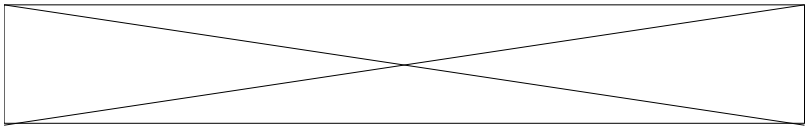
The probability theory has its origin in the games of chance pertaining to gambling. Jerome Cardon, an Italian mathematician was pioneer to write a book on 'Games of Chance' which was published posthumously in 1663. Credit goes to the French mathematicians Blaise Pascal and Pierre de Fermat who developed a systematic and scientific procedure for probability theory on solving the stake in an incomplete gambling match posed by a notable French gambler and noble man Chevalier de Mere. A Swiss mathematician James Bernoulli made an extensive study on probability is a major contribution to the theory of probability and combinatory. Contribution to the theory of probability was made by Abraham de Moivre (1667-1754) - The Doctrines of chances and the Revered Thomas Bayes (1702-61) - Inverse

Probability. In 19th century, Pierre Simon de Laplace made an extensive research on all the early ideas and published his monumental work under caption "Theory of Analytical Probability" in 1812 which became pioneer in theory of probability.

Meaning of Probability

The word probability is very commonly used in day-to-day conversation, and people have a rough idea about its meaning. We often come across statements like probably it may rain today. It means not sure about the occurrence of rain but there is possibility of occurrence of rain. What do we mean when we say the probability to win the match is 0.75? Do we mean that we will win three-fourths of the match. No. We actually mean that the conditions show the likelihood of winning the match is only 75 percent. Hence, the term probability is sensible numerical expression about uncertainty. In brief, probability is numerical value about uncertainty with calculated risk. If an event is certain not to occur, its probability is zero and if it is certain to occur, its probability is one.

The subject probability has been developed to a great extent and today, no discipline in social, physical or natural sciences is left without the use of probability. It is widely and popularly used in the quantitative analysis of business and economic problems; and is an essential tool in statistical inferences which form the basis for the decision theory. In other words, the role played by probability in modern science is of a substitute for certainty. Thus, the probability theory is a part of our everyday life.



Please use headphones

Terminology

The various terms which are used in defining probability under different approaches are discussed below.

a. Random Experiment:- An operation which produces an outcome is known as experiment. If it is conducted repeatedly under homogeneous conditions and the result is not unique but may be any one of the various possible outcomes, it is called random experiment. In other words, an experiment is said to be random if we cannot predict the outcome before the experiment is carried. For example, coin toss is a random experiment.
b. Trial and Event:- Performing a random experiment is called a trial, and the outcome is referred to as event. For example, the result is not unique by tossing a coin repeatedly. We may get either head or tail. Thus, tossing of a coin is a trial and getting head or tail is an event.

c. Independent and Dependent Events:- If the occurrence of an event does not affect the occurrence of other, such event is said to be independent event. For example, in tossing a die repeatedly, the event of getting number 3 in 1st throw is independent of getting number 3 in the 2nd, 3rd or subsequent throws. Event is said to be dependent, if the occurrence of an event affects the occurrence of the other. For example, when we draw a single card from a deck of cards, probability of it being a King is 4/52. If we do not replace this card, the probability of getting a King in the next draw will be affected (probability of getting a King in the second draw of card will be 3/51).
d. Mutually Exclusive Events :- Two or more events are said to be mutually exclusive if the occurrence of any one of them excludes the occurrence of all others. For example, if a die is thrown in a trial, by getting an outcome of say number 6, the occurrence of remaining numbers is excluded in that trial. Hence, all the outcomes are mutually exclusive.
e. Equally Likely Events:- The outcomes are said to be equally likely or equally probable if none of them is expected to occur in preference to other. For example, if an unbiased /fair coin is tossed, the outcomes heads and tails are equally likely to happen in each trial.
f. Exhaustive Events :- The total number of possible outcomes of a random experiment is called exhaustive events. For example, if a coin is tossed, we can get head (H) or tail (T). Hence exhaustive cases are 2. If two coins are tossed together, the various possible outcomes are HH, HT, TH, TT were HT means heads on the first coin and tails on the second coin and so on. Thus in a toss of two coins, exhaustive cases are 4, i.e. 2^2. In a throw of n coins, exhaustive events are 2^n.
g. Simple and Compound Events :- Events are said to be simple when it is about calculating probability of happening or not happening of single events. For example, the probability of drawing a red ball from a bag containing 8 white and 7 red balls. The joint occurrence of two or more events are termed compound events. For example if a bag containing 9 white and 6 red balls, and if two successive draws of 3 balls are made, we are going to find out the probability of getting 3 white balls in the first draw and 3 red balls in the second draw.
h. Complementary Events :- Let there be two events A and B. A is called complementary event of B (and vice-versa) if A and B are mutually exclusive and exhaustive events. For example, when a die is thrown, occurrence of an even number (2, 4, 6) and odd number (1, 3, 5) are complementary events.
i. Probability Tree:- A probability representation showing the possible outcomes of a series of experiments and their representative probabilities is known as Probability Tree.
j. Sample Space :- The set of all possible outcomes of a random experiment is known as the sample space and is denoted by S. In other words, the sample space is the set of all exhaustive cases of the random experiment. The elements of the sample space are the outcomes.

Algebra of Sets:

The union of two sets of A and B, denoted *by A* U *B* is defined as a: set of-elements which belong to either A or B or both. Symbolically,

AUB = (X ∈ A or X ∈ B)

For example, if

A = (1, 2, 3, 4} and B = {3, 4, 5, 6}

Then, AUB = (1, 2, 3, 4, 5, 6}

The intersection of two sets A and B, denoted by A(¯)B, is defined as a set of those elements that belong to both A and B. Symbolically,

A(¯)B = (X ∈ A and X ∈ B)

Thus in the above example,

A(¯)B = {3, 4}

Two sets A and B are said to be disjoint or mutually exclusive if they do not have any common elements. That is, A(¯)B = Ø.

In the figure given below, we represent the union, difference, and intersection of two events by means of Venn diagrams. The region enclosed by a rectangle is taken to represent the sample space whereas given events are represented by ovals within the rectangle.

Permutation and Combination

The word permutation means arrangement and the word combination means group or selection. For example, let us take three letters A, B and C. The permutations of these three letters taken two at a time will be AB, AC, BC, BA,CA and CB i.e. 6 in all, whereas the combinations of three letters taken two at a time will be AB, BC and CA i.e. 3 in all. The order of elements is immaterial in combinations, while in permutations the order of elements matters.

Permutation

A permutation of n different objects taken r at a time is an ordered arrangement of only r objects out of the n objects. In other words, the number of ways of arranging n things taken r at a time. It is denoted symbolically as ${}^{n}P_{r}$, where n is total number of elements in the set, r is the number of elements taken at a time, and P is the symbol for permutation. Thus,

$$ {}^{n}P_{r} = \frac{n!}{(n-r)!} $$

For example,

1. in how many ways can 4 books be arranged on a shelf?

Solution

We are given n=4, r=4. Number of possible ways of arrangement (or permutations) = 4! / (4-4)!

= 4!/1! = (4 x 3 x 2 x 1) / 1 (because 0! is 1)

= 24 permutations.

2. Find the number of permutations of the letters a, b, c, d, e taken 2 at a time.

Solution

n=5, r=2

${}^{n}P_r = {}^{5}P_2 = 5! / (5-2)! = 5! / 3! = (5 \times 4 \times 3 \times 2) / (3 \times 2 \times 1) = 20$ permutations

3. In how many ways can 8 differently coloured marbles be arranged in a row?

Solution

n=8, r=8

Number of ways we can arrange 8 differently coloured marbles are

${}^{n}P_r = {}^{8}P_8 = 8! / (8-8)! = 8! / 0! = (8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1) / 1 = 40320$

Combination

A combination is grouping or selection of all or part of a number of things without reference to the order of arrangement of the things selected. The number of combinations of n objects taken r at a time is denoted by ${}^{n}C_r$ given by

$$ {}^{n}C_r = \frac{n!}{(n-r)! \times r!} $$

4. In how many ways can a committee of 3 persons be chosen out of 8?

Solution

n=8, r=3

The possible number of ways of selection is given by

$^{n}C_r = {}^{8}C_3 = 8! / [(8-3)! \times 3!] = 8! / (5! \times 3!) = (8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1) / (5 \times 4 \times 3 \times 2 \times 1)(3 \times 2 \times 1)$

$= 8 \times 7 \times 6 / 3 \times 2 \times 1 = 56$ combinations

5. How many different sets of 4 students can be chosen out of 20 students?

Solution

n=20, r=4

The possible number of ways of selection is given by

$^{n}C_r = {}^{20}C4 = 20! / [(20-4)! \times 4!] = 20! / (16! \times 4!)$

$= (20 \times 19 \times 18 \times 17) / (4 \times 3 \times 2 \times 1) = 4845$ ways

6. In how many ways can a committee of 4 men and 3 women be selected out of 9 men and 6 women.

Solution

4 men can be selected from 9 men in $^{9}C_4$ ways = $9! / 5! \times 4! = (9 \times 8 \times 7 \times 6) / (4 \times 3 \times 2) = 126$ ways

3 women can be selected from 6 women in $^{6}C_3$ ways = $6! / 3! \times 3! = (6 \times 5 \times 4) / (3 \times 2 \times 1) = 20$ ways

So the number of ways of selecting 4 men from 9 and 3 women from 6 = $126 \times 20 = 2520$ ways.

Probability Theorems

The computation of probabilities can become easy and be facilitated to a great extent by the two fundamental theorems of probability - the Addition Theorem and the Multiplication Theorem.

(a) Addition Theorem - Independent Events:

The probability of occurring either event A or event B (where A and B are mutually exclusive events) is the sum of the individual probability of A and B. Symbolically,

$$P(A \text{ or } B) = P(A) + P(B)$$

Proof

If an event A can happen in a_1 ways and B in a_2 ways, then the number of ways in which either of the events can happen is $a_1 + a_2$. If the total number of possibilities is n, then by definition the probability of either A or B event happening is:

But,

and,

Hence P (A or B) = P(A) + P(B) - Proved

The theorem can be extended to three or more mutually exclusive events. Thus,

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$

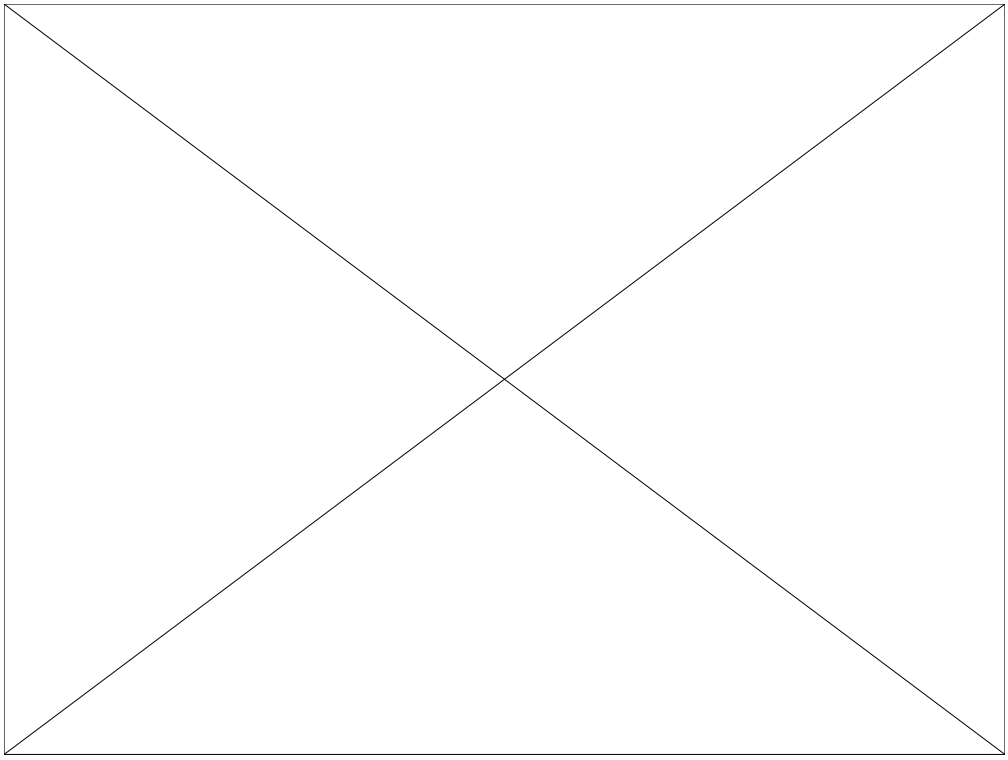
Addition Theorem - Dependent Events

If the events are not mutually exclusive, the above procedure discussed no longer holds. For example, if the probability of buying a pen is 0.6 and that of pencil is 0.3, we cannot calculate the probability of buying either pen or pencil by adding the two probabilities because the events are not mutually exclusive. When the events are not mutually exclusive, the above said theorem is to be modified. The probability of occurring of at least one of the two events, A and B which are not mutually exclusive is given by:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

By subtracting P(A and B) i.e. the proportion of events as counted twice in P(A) + P(B), the addition theorem is, thus, reconstructed in such a way as to render A and B mutually exclusive events. In case of three events :

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) - P(A \text{ and } B) - P(A \text{ and } C) - P(B \text{ and } C) + P(A \text{ and } B \text{ and } C)$$



Please use headphones

Illustration 1

A bag contains 25 balls marked 1 to 25. One ball is drawn at random. What is the probability that it is marked with a number that is multiple of 5 or 7?

Solution

The sample space i.e. the possible outcomes = 25.

i. The possible sample points that are multiples of 5 are 5 ...(5, 10, 15, 20, 25)

So probability of getting a multiple of 5 = 5/25

ii. The possible sample points that are multiples of 7 are 3 ...(7, 14, 21)

So probability of getting a multiple of 7 = 3/25

So the probability that the ball picked at random is either marked with a multiple of 5 or 7

= 5/25 + 3/25 = 8/25

Illustration 2

A bag contains 20 balls numbered 1 to 20. One ball is drawn at random. What is the probability that it is marked with a number that is a multiple of 3 or 4?

Solution

The possible outcomes = 20.

i. The possible outcomes where the ball with a multiple of 3 = 6 ...(3, 6, 9, 12, 15, 18)

So probability of getting a multiple of 3 = 6/20

ii. The possible outcomes where the ball has a multiple of 4 = 5 ...(4, 8, 12, 16, 20)

So probability of getting a multiple of 4 = 5/20

So the probability that the ball picked at random is either marked with a multiple of 3 or 4 is not just adding the two probabilities as calculated above. Since there are some numbers which are multiples of both 3 and 4, we need to exclude them so that they don't get counted twice.

iii. The possible outcomes where the ball has a multiple of both 3 and 4 = 1 ...(12)

Hence, the probability of two events which are not mutually exclusive, i.e., disjoint cases = 1/20

Probability of drawing either a multiple of 3 or 4 = 6/20 + 5/20 - 1/20 = 10/20 = 1/2

Illustration 3

From a pack of 52 cards, what is the probability of drawing one card that it is either a King or a Queen?

Solution

Probability of drawing a King = P(K) = 4/52 (as there are 4 Kings in a pack of 52 cards) = 1/13

Probability of drawing a Queen = P(Q) = 4/52 (as there are 4 Queens in a pack of 52 cards) = 1/13

Probability of getting either a King or Queen in the draw = P(K or Q) = P(K) + P(Q) = 1/13 + 1/13 = 2/13

Illustration 4

A bag contains 30 balls numbered from 1 to 30. One ball is drawn at random. Find the probability that the number of

drawn ball will be a multiple of

(a) 5 or 9, and

(b) 5 or 7

Solution

a) Probability of getting a ball that has a multiple of 5, i.e. 5, 10, 15, 20, 25 or 30 = 6/30

Probability of getting a ball that has a multiple of 9, i.e. 9, 18, 27 = 3/30

Then, the probability of getting a ball that has either multiple of 5 or 9 = 6/30 + 3/30 = 9/30 = 3/10

b) Probability of getting a ball that has a multiple of 5, i.e. 5, 10, 15, 20, 25 or 30 = 6/30

Probability of getting a ball that has a multiple of 7, i.e. 7, 14, 21, 28 = 4/30

Then, the probability of getting a ball that has either multiple of 5 or 7 = 6/30 + 4/30 = 10/30 = 1/3

Illustration 5

The probability that a student passes a Chemistry test is 2/3 and the probability that he passes both Chemistry and English tests is 14/45. The probability that he passes at least one test is 4/5. What is the probability that he passes the English test?

Solution

Probability that the student passes a Chemistry test = P(C) = 2/3

Probability that the student passes an English test = P(E) = ?

Probability that the student passes both Chemistry and English tests = P(C and E) = 14/45

Probability that the student passes either Chemistry or English test = P(C or E) = 4/5

Using the Addition Theorem of Probability - Dependent Events,

$$P(C \text{ or } E) = P(C) + P(E) - P(C \text{ and } E)$$

Putting values in the formula, we get

4/5 = 2/3 + P(E) - 14/45

P(E) = 4/5 + 14/45 - 2/3 = (36+14)/45 = 50/45 - 2/3 = 10/9 - 2/3 = (10-6)/9 = 4/9

Illustration 6

Let A and B be the two possible outcomes of an experiment and suppose that P(A) = 0.4, P(A or B) = 0.8,

and P(B) = P. Then,

i. For what choice of P are A and B mutually exclusive?

ii. For what choice of P are A and B independent?

Solution

i. We know that

P(A or B) = P(A) + P(B) – P(A and B)

P(A and B) = P(A) + P(B) – P(A or B)

= 0.4 + P – 0.8

= P - 0.4

If A and B are to be mutually exclusive, then there should be no outcomes common to both.

That is, P(A and B) = 0

⇒ P - 0.4 = 0

⇒ P = 0.4

ii. If A and B are to be independent, then P(A and B) should be equal to P(A) x P(B)

⇒ P – 0.4 = 0.4 x P

⇒ P - 0.4P = 0.4

⇒ 0.6P = 0.4

⇒ P = 0.4 / 0.6 = 2/3

⇒ P = 0.677

Illustration 7

A person is known to hit the target in 3 out of 4 shots, whereas another person is known to hit the target in 2 out of 3 shots. Find the probability of the target being hit at all when they both try.

Solution

Probability that the first person hits the target = 3/4

Probability that the second person hits the target = 2/3

Since the events are not mutually exclusive,

P(A or B) = P(A) + P(B) – P(A and B) = P(A) + P(B) - P(A) x P(B)

= 3/4 + 2/3 - (3/4) x (2/3) = (9+8)/12 - 1/2 = 17/12 - 6/12 = 11/12

Illustration 8

a. A product is assembled from three components - X, Y, and Z. The probability of these components being defective is 0.01, 0.02 and 0.05 respectively. What is the probability that the assembled product will not be defective?
b. Items produced by a certain process may have one or both of the two types of defects A and B. It is known that 20 per cent of the items have type A defects and 10 per cent have type B defects. Furthermore, 6 per cent are known to have both types of defects. What is the probability that a randomly selected item will be defective?

Solution

a. Let P(X), P(Y) and P(Z) be the respective probabilities of components of X, Y and Z being defective. We are given that P(X) = 0.01, P(Y) = 0.02, P(Z) = 0.05.

   The probability that the assembled product will be defective = P(X or Y or Y)

   = P(X) + P(Y) + P(Z) - P(XY) - P(YZ) - P(XZ) + P(XYZ)

   = 0.01 + 0.02 + 0.05 - 0.01 x 0.02 - 0.02 x 0.05 + 0.01 x 0.02 x 0.05

   = 0.08 - 0.0002 - 0.0010 + 0.000010

   = 0.080010 - 0.001200 = 0.078810

   The probability that the assembled product will not be defective = 1 - Probability that it will be defective

   = 1 - 0.078810 = 0.921190 = 0.92

b. Let P(A) be the probability that an item will have A-type of defect, and P(B) be the probability that an item will have B-type of defect. We are given that P(A) = 20% = 0.2, P(B) = 10% = 0.1, P(A and B) = 6% = 0.06.

   Probability that a randomly selected item will be defective is P(A or B)

   = P(A) + P(B) - P(A and B)

= 0.2 + 0.1 - 0.06 = 0.24

Illustration 9

A salesman has 65 per cent chance of making a sale to a customer. The behaviour of each successive customer is independent. If three customers A, B and C enter together, what is the probability that the salesman will make a sale to at least one of the customers.

Solution

Let the probability of making sale to A = P(A), probability of making sale to B = P(B), and probability of making sale to C = P(C)

P(A) = P(B) = P(C) = 0.65

Probability that sale is made to at least one of the customers = P(A or B or C)

= P(A) + P(B) + P(C) – P(AB) - P(BC) - P(AC) + P(ABC)

= 0.65 + 0.65 + 0.65 - 0.65 x 0.65 - 0.65 x 0.65 – 0.65 x 0.65 + 0.65 x 0.65 x 0.65

= $3 \times 0.65 - 3 \times (0.65)^2 + (0.65)^3$

= 1.95 - 1.2675 + 0.274625 = 0.957125 = 0.957

(b) Multiplication Theorem:

The probability of occurring of two independent events A and B is equal to the product of their individual probabilities.

Symbolically, if A and B are independent events, then

P(A and B) = P(A) x P(B)

Proof

If an event A can happen in n_1 ways of which a_1 are successful, we can combine each successful event in the first with each successful event in the second. Thus, the total number of successful events in both cases is $a_1 \times a_2$. Similarly, the total number of possible cases is $n_1 \times n_2$. By definition, the probability of occurrence of both events is:

$$\frac{a_1 \times a_2}{n_1 \times n_2} = \frac{a_1}{n_1} \times \frac{a_2}{n_2}$$

Therefore, P(A and B) = P(A) x P(B).

Illustration 10

Four cards are drawn at random from a pack of 52 cards. Find the probability that

i. They are a King, a Queen, a Jack and an Ace

ii. Two are Kings and two are Aces

iii. All are Diamonds

iv. Two are red and two are black

v. There is one card of each suit

vi. There are two cards of Clubs and two cards of Diamonds

Solution

A pack contains 52 cards. We can draw 4 cards from 52 cards, in ${}^{n}C_{r}$ i.e. ${}^{52}C_{4}$ ways which gives exhaustive number of cases i.e. the sample space.

i. A pack of 52 cards consisting of four cards each of King, Queen, Jack and Ace. So, a King or a Queen or a Jack or an Ace can be drawn in ${}^{4}C_{1} = 4$ ways. Since drawing a King is independent of the ways of getting a Queen, a Jack and an Ace, the sample points or the favorable number of cases are ${}^{4}C_{1} \times {}^{4}C_{1} \times {}^{4}C_{1} \times {}^{4}C_{1}$.

Probability = No. of favourable cases / Total number of exhaustive cases

$= ({}^{4}C_{1} \times {}^{4}C_{1} \times {}^{4}C_{1} \times {}^{4}C_{1}) / {}^{52}C_{4}$

$= 256 / 270725 = 0.000945$

ii. Probability of getting two Kings and two Queens $= ({}^{4}C_{2} \times {}^{4}C_{2}) / {}^{52}C_{4} = 36 / 270725 = 0.00013$

iii. A pack of 51 cards contains 13 Diamond cards. So we can draw 4 out of 13 Diamond cards in ${}^{13}C_{4}$ ways.

Probability of getting all diamonds $= {}^{13}C_{4} / {}^{52}C_{4} = 0.00264$

iv. Probability of getting two reds and two blacks $= ({}^{26}C_{2} \times {}^{26}C_{2}) / {}^{52}C_{4} = 0.390$

v. In a pack of 52 cards, there are 13 cards of each suit. We can draw one card of each suit in ${}^{13}C_{1} \times {}^{13}C_{1} \times {}^{13}C_{1} \times {}^{13}C_{1}$ ways.

Probability of getting one card of each suit $= ({}^{13}C_{1} \times {}^{13}C_{1} \times {}^{13}C_{1} \times {}^{13}C_{1}) / {}^{52}C_{4} = 0.1055$

vi. In a pack of 52 cards, there are 13 Club cards and 13 Diamond cards. So, we can draw two cards of Clubs and two cards of Diamonds in ${}^{13}C_{2} \times {}^{13}C_{2}$ ways.

Probability of getting two cards of Diamonds and two cards of Clubs $= ({}^{13}C_{2} \times {}^{13}C_{2}) / {}^{52}C_{4} = 0.0225$

Illustration 11

A bag contains 8 white and 6 red balls. Two balls are drawn in succession at random. What is the probability that one of them is white and the other is red?

Solution

The sample space or exhaustive outcomes = $^{14}C_2 = 91$

One white ball out of eight can be drawn in $^{8}C_1 = 8$ ways

One red ball out of six can be drawn in $^{6}C_1 = 6$ ways

Number of favourable cases for both white and red balls = 8 x 6 = 48 ways

Probability of getting one white and one red ball = 48 / 91 = 0.527

Illustration 12

An urn contains 8 white and 3 red balls. If two balls are drawn at random, find the probability that

(i) both are white

(ii) both are red

(iii) one is of each colour

Solution

Total balls in the urn = 11

2 balls drawn at a time, can be drawn in $^{11}C_2$ ways = 55 ways

(i) 2 white balls can be drawn out of 8 in $^{8}C_2$ ways = 28 ways

Probability that both balls are white = 28/55

(ii) 2 red balls can be drawn out of 3 in $^{3}C_2$ ways = 3 ways

Probability that both balls are red = 3/55

(iii) 1 white ball and 1 red ball can be drawn out of 8 white and 3 red balls in $^{8}C_1$ x $^{3}C_1$ways = 24 ways

Probability that one ball is of each colour = 24 / 55

Illustration 13

A bag contains 10 white, 6 red, 4 black and 7 blue balls. 5 balls are drawn at random. What is the probability that 2 are red, 2 are white, and 1 is blue?

Solution

Total balls = 27

No. of balls drawn = 5

5 balls can be drawn out of 27 balls in $^{27}C_5$ ways = 80730 ways

2 red balls can be drawn out of 6 red balls in $^{6}C_2$ ways = 15 ways

2 white balls can be drawn out of 10 white balls in $^{10}C_2$ ways = 45 ways

1 blue ball can be drawn out of 7 blue balls in $^{7}C_1$ ways = 7 ways

Probability of getting 2 red balls, 2 white balls and 1 blue ball = (15 x 45 x 7) / 80730 = 0.585

Illustration 14

A bag contains 8 red and 5 white balls. Two successive drawings of 3 balls are made. Find the probability that the first drawing will give 3 white and the second 3 red balls, if

(i) the balls were replaced before the second trial,

(ii) the balls were not replaced before the second trial.

Solution

Let us define the events,

A = Drawing 3 white balls in the first draw

B = Drawing 3 red balls in the second draw

P(AB) = Probability of drawing 3 white and 3 red balls in two draws.

(i) Since the balls were replaced before the second draw, events A and B are independent events.

In first draw, 3 balls out of 13 can be drawn in $^{13}C_3$ ways (286 ways) which is the exhaustive number of cases.

3 white balls out of 5 white balls can be drawn in $^{5}C_3$ ways (10 ways).

P(A) = $^{5}C_3$ / $^{13}C_3$ = 10/286

In second draw, 3 balls out of 13 can be drawn in $^{13}C_3$ ways (286 ways) which is the exhaustive number of cases.

3 red balls out of 8 red balls can be drawn in $^{8}C_3$ ways (56 ways).

$P(B) = {}^{8}C_3 / {}^{13}C_3 = 56/286$

Hence, probability of getting 3 white and 3 red balls $P(AB) = P(A) \times P(B) = (10/286) \times (56/286) = 0.006846$

ii. Since the balls were not replaced before the second draw, events A and B are dependent events.

In first draw, 3 balls out of 13 can be drawn in $^{13}C_3$ ways (286 ways) which is the exhaustive number of cases.

3 white balls out of 5 white balls can be drawn in $^{5}C_3$ ways (10 ways).

$P(A) = {}^{5}C_3 / {}^{13}C_3 = 10/286$

In second draw, 3 balls out of 10 can be drawn in $^{10}C_3$ ways (120 ways) which is the exhaustive number of cases.

3 red balls out of 8 red balls can be drawn in $^{8}C_3$ ways (56 ways).

$P(B) = {}^{8}C_3 / {}^{10}C_3 = 56/120$

Hence, probability of getting 3 white and 3 red balls $P(AB) = P(A) \times P(B) = (10/286) \times (56/120) = 0.01632$

Illustration 15

An urn contains 5 white, 3 black, and 6 red balls. 3 balls are drawn at random. Find the probability that:

i. two of the balls drawn are white

ii. each one is of different colour

iii. none is black

iv. at least one is white

Solution

Total balls is urn are 14. 3 balls can be drawn out of 14 balls in $^{14}C_3$ ways, i.e., the exhaustive number of cases = 364.

i. 2 white balls can be drawn out of 5 in $^{5}C_2$ ways and another ball can be drawn from the remaining 9 balls in $^{9}C_1$ ways. So, the favourable cases are $^{5}C_2 \times {}^{9}C_1 = 90$.

Required probability = 90 / 364 = 0.247

ii. We can draw 3 different coloured balls in ${}^5C_1 \times {}^3C_1 \times {}^6C_1 = 90$ ways.

Required probability = 90 / 364 = 0.247

iii. We can draw 3 balls out of the 11 non-black balls in ${}^{11}C_3 = 165$ ways.

Required probability = 165 / 364 = 0.453

iv. P(at least 1 white ball) = 1 - P(no white ball).

We can draw 3 balls out of 9 non-white balls in ${}^9C_3 = 84$ ways.

Probability of drawing 3 non-white balls or in other words no white ball = 84 / 364 = 0.231

Probability of drawing at least one white ball = 1 - 0.231 = 0.769

Illustration 16

The probability that India wins a cricket test match against Australia is given to be 1/3. If India and Australia play 4 test matches, what is the probability that

(i) India will lose all the four test matches,

(ii) India will win at least one test match.

Solution

In notation,

Let event A = India wins

and event B = India loses

Probability(India wins 1st test match) = P(A1)

Probability(India wins 2nd test match) = P(A2)

Probability(India wins 3rd test match) = P(A3)

Probability(India wins 4th test match) = P(A4)

P(A1) = P(A2) = P(A3) = P(A4) = 1/3

Probability(India loses 1st test match) = P(B1)

Probability(India loses 2nd test match) = P(B2)

Probability(India loses 3rd test match) = P(B3)

Probability(India loses 4th test match) = P(B4)

P(B1) = P(B2) = P(B3) = P(B4) = 2/3

(i) Probability that India will lose all 4 test matches = 2/3 x 2/3 x 2/3 x 2/3 = 16/81 = 0.19

(ii) Probability that India will win at least one test match

= 1 - Probability that India will lose all 4 test matches = 1 - 0.19 = 0.81

Illustration 17

A University has to select an examiner from a list of 60 persons, 25 of them women and 35 men; 15 of them knowing Hindi and 45 not, 18 of them being teachers and the remaining 42 not. What is the probability of the University selecting a Hindi-knowing woman teacher?

Solution

Probability of the selected person being a woman = P(W) = 25/60

Probability of the selected person knowing Hindi = P(H) = 15/60

Probability of the selected person being a teacher = P(T) = 18/60

Probability of the selected person being a Hindi-knowing woman teacher

= P(WHT) = 25/60 x 15/60 x 18/60 = 0.03125

Illustration 18

The MCom class consists of 60 students, 12 of them are girls and 48 boys, 10 of them are rich, 15 of them are fair complexioned. What is the probability of selecting a fair complexioned rich girl?

Solution

Probability of the selected person being a girl = P(G) = 12/60

Probability of the selected person being rich = P(R) = 10/60

Probability of the selected person being fair-complexioned = P(F) = 15/60

Probability of the selected person being a fair complexioned rich girl

= P(GRF) = 12/60 x 10/60 x 15/60 = 0.00833

Illustration 19

The following table shows Subjects and Educational qualification of 40 teachers of S.K. University. Of total 144, one is nominated at random to be the University Executive. Find the probability that

(i) the teacher has only a PG degree,

(ii) the teacher has a PhD degree and is from Commerce subject,

(iii) the teacher has a PhD degree and if from Economics subject,

(iv) the teacher has a PhD degree.

The data is:

Subject	PG Degree only	PhD Degree
Commerce	3	9
Economics	5	12
Rural Development	4	7

Solution

Total number of teachers = 40

(i) Probability of nominating a teacher with only PG degree

= 12 / 40 = 0.3

(ii) Probability of nominating a teacher with PhD degree and Commerce subject

= 9/40 = 0.225

(iii) Probability of nominating a teacher with PhD degree and Economics subject

= 12/40 = 0.3

(iv) Probability of nominating a teacher with PhD degree

= 9/40 x 12/40 x 7/40 = 0.0118.

- End of Chapter -

LESSON - 6

CONDITIONAL EVENTS

Conditional Theorem:

If two events A and B are dependent, the probability of the second event occurring will be affected by the outcome of the first that has already occurred. The term conditional probability is used to describe this situation. It is symbolically denoted by P(B/A), which is read as the probability of occurring B, given that A has already occurred. Robert L. Birte defined the concept of conditional probability as: "A conditional probability indicates that the probability that an event will occur is subject to the condition that another event has already occurred."

Symbolically,

$P(B/A) = P(AB) / P(A)$

and

$P(A/B) = P(AB) / P(B)$

In conditional probabilities, the rule of multiplication in its modified form is:

$P(A \text{ and } B) = P(A) \times P(B/A) = P(B) \times P(A/B)$

Bayes' Rule

Computation of unknown probabilities on the basis of information supplied by the past records, or experiment is one of the most important applications of the conditional probability. The occurrence of an event B only, when event A is known to have occurred (or vice versa) is said to be conditional probability, which is denoted by P(B/A). In other words, probability of B, given A. The conditional probability occuring due to a particular event or reason is called its reverse or posteriori probability. The posteriori probability is computed by Bayes' Rule, named after its innovator, the British Mathematician, Sir Thomas Bayes. The revision of given i.e. old probabilities in the light of the additional information supplied by the experiment or past records is of extreme help to business and management executives in making valid decisions in the face of uncertainties. The Bayes' theorem is defined as:

Probability of event A1 given event B, $P(A1/B) = P(A1 \text{ and } B) / P(B)$

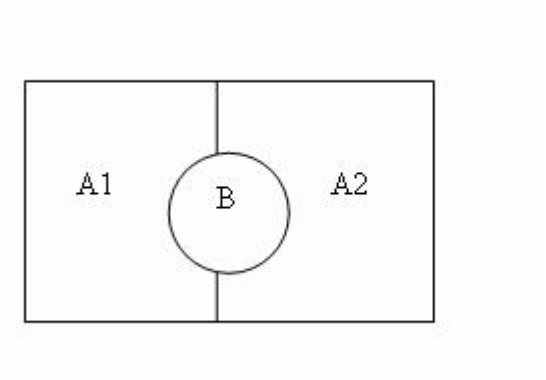
Probability of event A2 given event B, P(A2/B) = P(A2 and B) / P(B)

where P(B) = P(A1 and B) + P(A2 and B)

$$P(A_1 \text{ and } B) = P(A_1) \times P\left(\frac{B}{A_1}\right)$$

$$P(A_2 \text{ and } B) = P(A_2) \times P\left(\frac{B}{A_2}\right)$$

By observing the following diagram, the derivation of the above formula can be drawn



A_1 and A_2 = The set of events which are mutually exclusive

B = An event which intersects each of the events

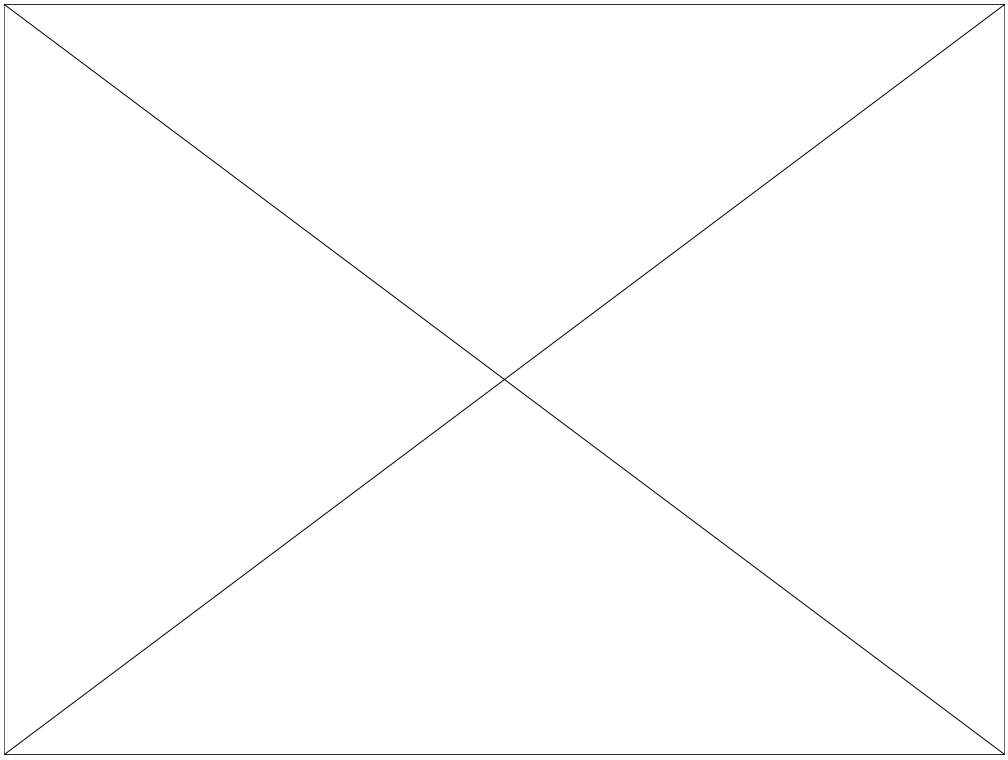
(Observe diagram. The part of B which is within A1 represents the area (A1 and B); and the part of B which is within A2 represents the area (A2 and B).

F.R. Jolliffee in his book captioned 'Commonsense Statistics for Economists and Others' has defined the concept of Bayes theorem as: With twisting conditional probabilities the other way round, i.e. given probabilities of the form P(B/Ai) to find conditional probabilities of the form P(Ai/B), where Ai can be described as prior probabilities i.e. probabilities known before anything happens, then the probabilities P(Ai/B) are described as posterior probabilities i.e. probabilities found after something has happened, and supported experimental evidences.

Probability before revision by Bayes' rule is called a priori or simply prior probability, since it is determined before the sample information is taken in account. A probability which has undergone revision via Bayes' rule is called posterior probability because, it represents a probability computed after the sample information is taken into account.

Posterior probability is also called revised probability in the sense that it is obtained by revising the prior probability with the sample information. Posterior

probability is always conditional probability, the conditional event being the sample information. Thus, a prior probability which is unconditional probability becomes a posterior probability, which is conditional probability by using Baye's rule.



Please use headphones

Illustration 20

Two sets of candidates are competing for positions on the Board of Directors of a company. The probability that the first and second sets will win are 0.65 and 0.35 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 0.3. What is the probability that the product will be introduced.

Solution

Let A_1 = event of the first set of candidates getting selected

A_2 = event of the second set of candidates getting selected

B = event of introducing a new product.

We are given that,

$P(A_1) = 0.65$

$P(A_2) = 0.35$

$P(B/A_1) = 0.80$

$P(B/A_2) = 0.30$

Probability that the product will be introduced, P(B)

$= P(A_1 \text{ and } B) + P(A_2 \text{ and } B)$

$= P(A1) \times P(B/A1) + P(A2) \times P(B/A2)$

$= 0.65 \times 0.80 + 0.35 \times 0.30$

$= 0.52 + 0.105$

$= 0.625 = 62.5\%$

Illustration 21

A factory has two machines. Past records show that the first machine produces 40 per cent of output and the second machine produces 60 per cent of output. Further, 4 per cent and 2 per cent of products produced by the first machine and the second machine respectively, were defectives. If a defective item is drawn at random, what is the probability that the defective item was produced by the first machine or the second machine.

Solution

Let A_1 = the event of drawing an item produced by the first machine

A_2 = the event of drawing an item produced by the second machine

B = the event of drawing a defective item produced by these machines

Based on the information given,

P(A1) = 40% = 0.4

P(A2) = 60% = 0.60

P(B/A1) = 4% = 0.04

P(B/A2) = 2% = 0.02

P(B) = P(A1 and B) + P(A2 and B)

$= P(A1) \times P(B/A1) + P(A2) \times P(B/A2)$

$= 0.40 \times 0.04 + 0.60 \times 0.02$

= 0.016 + 0.012 = 0.028

P(A1/B) = P(A1 and B) / P(B) = P(A1) x P(B/A1) / P(B) = 0.4 x 0.04 / 0.028 = 0.571

P(A2/B) = P(A2 and B) / P(B) = P(A2) x P(B/A2) / P(B) = 0.6 x 0.02 / 0.028 = 0.429

The values derived above can also be computed in a tabular form as below.

Computation of posterior probabilities

Event	Prior Probability	Conditional probability	Joint probability	Posterior probability	
	P(A1)	P(B/A1)	P(A1 and B)	P(A1/B)	
(1)	(2)	(3)	(4)	(5) = (4)/P(B)	
A_1	0.40	0.04	0.016	0.016/0.028	= 0.5714
A2	0.60	0.02	0.012	0.012/0.028	= 0.4286
	1.00		P(B) = 0.028	1.0000	

Conclusion

With additional information i.e. conditional probability, the probability of defective items produced by the first machine is 0.5714 or 57.14 per cent and that by the second machine is 0.4286 or 42.86 per cent. And we may say that the defective item is more likely drawn from the output produced by the first machine.

Illustration 22

The probability that management trainee will remain with a company is 0.60. The probability that an employee earns more than Rs. 10,000 per year is 0.50. The probability that an employee is a management trainee who remained with the company or who earns more than Rs. 10000 per year is 0.70. What is the probability that an employee earns more than Rs. 10,000 per year given that he is a management trainee who stayed with the company.

Solution

Let us define the events,

A = Management trainee will remain with the company

B = An employee earns more than Rs. 10,000.

P(A) = 0.60

P(B) = 0.50

P(A or B) = 0.70

P(A or B) = P(A) + P(B) – P(AB)

0.70 = 0.60 + 0.5 – P(AB)

P(AB) = 0.4

P(B/A) = P(AB) / P(A) = 0.4 / 0.6 = 0.67

Probability of an employee earns more than Rs. 10000 per year given that he is a management trainee is 0.67.

Illustration 23

Suppose that one of the three men, a politician, a businessman and an academician will be appointed as the Vice Chancellor of a University. Their probability of appointments respectively are 0.40, 0.25 and 0.35. The probabilities that research activities will be promoted by these people if they are appointed are 0.5, 0.4 and 0.8 respectively. What is the probability that research will be promoted by the new Vice-Chancellor?

Solution

Let us define the events,

A_1 = Event that the politician will be appointed as the Vice-Chancellor

A_2 = Event that the businessman will be appointed as the Vice-Chancellor

A_3 = Event that the academician will be appointed as the Vice-Chancellor

B = Event that research activities will be promoted

Given that $P(A_1) = 0.40$, $P(A_2) = 0.25$, $P(A_2) = 0.35$

And, $P(B/A_1) = 0.50$, $P(B/A_2) = 0.40$, $P(B/A_3) = 0.80$

The probability that research will be promoted by the new Vice-Chancellor, P(B)

$= P(A_1) \times P(B/A_1) + P(A_2) \times P(B/A_2) + P(A_3) \times P(B/A_3)$

$= 0.40 \times 0.50 + 0.25 \times 0.40 + 0.35 \times 0.80$

$= 0.20 + 0.10 + 0.28 = 0.58$

Now,

Hence, the appointment of a politician, a businessman and an academician as the Vice-Chancellor will promote research is probable 0.3448 or 34.48%, 0.1724 or 17.24%, and 0.4828 or 48.28% respectively. Appointment of an academician as the Vice-Chancellor would certainly develop and promote education as it is known by the probability theory.

Random Variable and Probability Distribution

By random variable we mean a variable value which is computed by the outcome of a random experiment. In brief, a random variable is a function which assigns a unique value to each sample point of the sample space. A random variable is also called a chance variable or stochastic variable. A random variable may be continuous or discrete. If the random variable takes on all values within a certain interval, then it is called a continuous random variable while if the random variable takes on the integer values such as 0, 1, 2, 3, ... then it is known as discrete random variable.

The function p(x) is known as the probability function of random variable of x and the set of all possible ordered pairs is called probability distribution of random variable. The concept of probability distribution is in relation to that of frequency distribution. While the frequency distribution tells how the total frequency is distributed among different classes of the variable, the probability distribution tells how the total frequency is distributed among different classes of the variable, the probability distribution tells how the total probability of 1 is distributed among various values which random variable can take. In brief, the word frequency is replacing by probability.

Illustration 24

A dealer of Allwyn refrigerators estimates from his past experience the probabilities of his selling refrigerators in a day. These are as follows:

No. of refrigerators	:	0	1	2	3	4	5	6
Probability	:	0.03	0.20	0.23	0.25	0.12	0.10	0.07

Find the mean number of refrigerators sold in a day.

Solution

$P(X) = P(X = Xi)$, where $i = 1, 2, 3 \ldots n$

$P(X) = 0 \times 0.03 + 1 \times 0.20 + 2 \times 0.23 + 3 \times 0.25 + 4 \times 0.12 + 5 \times 0.10 + 6 \times 0.07$

$= 0 + 0.20 + 0.46 + 0.75 + 0.48 + 0.50 + 0.42$

$= 2.81$

Hence, the mean number of refrigerators sold in a day is 2.81

Illustration 25

A die is tossed twice. Getting an odd number is termed as a success. Find the probability distribution of number of success.

Solution

Getting an odd number in a throw of a die has 3 possibilities in all - 1, 3, or 5

So, probability of success = 3/6 = 1/2

Probability of failure = 1 probability of success = 1 - 1/2 = 1/2

Let us denote success by S and failure by F. In two throws of a die, X denoted by S becomes a random variable and takes the values 0, 1, 2. This means, in the two throws, we can get either no odd numbers, or 1 odd number, or both odd numbers.

i. P(X=0) = P(F in both throws) = P(FF) = P(F) x P(F) = 1/2 x 1/2 = 1/4

ii. P(X=1) = P(S in first throw and F in second throw) or P(F in first throw and S in second throw) = P(SF) or P(FS)

= P(S) x P(F) + P(F) x P(S) = 1/2 x 1/2 + 1/2 x 1/2 = 1/4 + 1/4 = 2/4 = 1/2

iii. P(X=2) = P(S in both throws) = P(SS) = P(S) x P(S) = 1/2 x 1/2 = 1/4

Thus, the probability distribution of X is :

X	0	1	2
P(X)	0.25	0.50	0.25

Illustration 26

Four bad apples are mixed accidentally with 26 good apples. Obtain the probability distribution of the number of bad apples in a draw of 3 apples at random.

Solution

Denote X as the number of bad apples drawn. Now X is a random variable which takes values of 0, 1, 2, 3. There are 30 apples in all (26 good + 4 bad), and the exhaustive number of cases drawing three apples is ${}^{30}C_3$. We get,

P(X=0) = 0 bad apples and 3 good apples = $({}^{4}C_0 \times {}^{26}C_3) / {}^{30}C_3$ = 1 x 26 x 25 x 24 / 30 x 29 x 28 = 2600 /4060

P(X=1) = 1 bad apple and 2 good apples = $({}^{4}C_1 \times {}^{26}C_2) / {}^{30}C_3$ = 4 x 26 x 25 x 6 / 2 x 30 x 29 x 28 = 1300 / 4060

P(X=2) = 2 bad apples and 1 good apple = $({}^4C_2 \times {}^{26}C_1) / {}^{30}C_3 = 4 \times 3 \times 26 \times 6 / 2 \times 30 \times 29 \times 28 = 156 / 4060$

P(X=3) = 3 bad apples and 0 good apples = $({}^4C_3 \times {}^{26}C_0) / {}^{30}C_3 = 4 \times 1 \times 6 / 30 \times 29 \times 28 = 4 / 4060$

Hence, the probability distribution of X is:

X	0	1	2	3
P(X)	2600/4060	1300/4060	156/4060	4/4060

Illustration 27

An insurance company offers a 40 year old man a Rs. 1200 one year term insurance policy for an annual premium of Rs. 15. Assume that the number of deaths per one thousand is four persons in this group. What is the expected gain for the insurance company on a policy of this type?

Solution

Denote premium by X and death rate by P(X).

Accordingly,

P(X) = 4/1000 = 0.004

Probability of no death = 1 - P(X) = 1 - 0.004 = 0.996

Expected gain for the insurance company, E(X)

= Premium charged x Probability of no death - Net amount payable on death x Probability of death

= 15 x 0.996 - 1185 x 0.004 = 14.94 - 4.74 = Rs.10.20

Illustration 28

Gopi, a newspaper vendor purchases the Hindu newspaper at a special concessional rate of Rs. 1.80 per copy against the selling price of Rs. 2.30. Any unsold copies are, however, a dead loss. Gopi has estimated the following probability distribution for the number of copies demanded.

No. of copies :	20	21	22	23	24	25
Probability :	0.05	0.18	0.30	0.28	0.10	0.08

How many copies should he order so that the expected profit is the maximum.

Solution

Profit per copy = 2.30 - 1.80 = Rs. 0.50

Expected profit = Profit per copy x Probability of paper distribution x No. of papers sold.

No. of copies	Probability	Profit per copy	Expected profit
20	0.05	0.50	20 x 0.05 x 0.5 = 0.50
21	0.18	0.50	21 x 0.18 x 0.5 = 1.89
22	0.30	0.50	22 x 0.30 x 0.5 = 3.30
23	0.28	0.50	23 x 0.28 x 0.5 = 3.22
24	0.10	0.50	24 x 0.10 x 0.5 = 1.20
25	0.08	0.50	25 x 0.08 x 0.5 = 1.00

By purchasing and selling 22 copies of the newspaper, Gopi will get maximum expected profit of Rs. 3.30

Illustration 29

(a) The monthly demand for Allwyn watches is known to have the following probability distribution:

Demand	:	1	2	3	4	5	6	7	8
Probability	:	0.08	0.12	0.19	0.24	0.16	0.10	0.07	0.04

Determine the expected demand for watches. Also compute the variance.

(b) A random variable has the following probability distribution.

X	-1	0	1	2
P(X)	1/3	1/6	1/6	1/3

Compute the expectation of X.

Solution

a) Expected demand of Allwyn watches = Xi x Pi

$E(X) = 1 \times 0.08 + 2 \times 0.12 + 3 \times 0.19 + 4 \times 0.24 + 5 \times 0.16 + 6 \times 0.10 + 7 \times 0.07 + 8 \times 0.04$

$= 0.08 + 0.24 + 0.57 + 0.96 + 0.80 + 0.60 + 0.49 + 0.32 = 4.06$

$E(X)^2 = Xi^2 P(Xi)$

$= 1^2 \times 0.08 + 2^2 \times 0.12 + 3^2 \times 0.19 + 4^2 \times 0.24 + 5^2 \times 0.16 + 6^2 \times 0.10 + 7^2 \times 0.07 + 8^2 \times 0.04$

$= 0.08 + 0.48 + 1.71 + 3.84 + 4.00 + 3.60 + 3.43 + 2.56 = 19.7$

Variance = $E(X)^2 - [E(X)]^2 = 19.7 - (4.06)^2 = 3.22$

b) Expectation of X = Xi x P(Xi) = $-1 \times 1/3 + 0 \times 1/6 + 1 \times 1/6 + 2 \times 1/3 = -1/3 + 1/6 + 2/3 = 3/6 = 0.50$

Illustration 30

Ravali, proprietor of a food stall has invented a new item of food delicacy which she calls R-foods. She has calculated that the cost of manufacturing as Rs. 2 per piece, and because of its quality it would the sold for Rs. 3 per piece. It is, however, perishable and any goods unsold at the end of the day are dead loss. She expects the demand to be variable and has drawn up the following probability distribution:

No. of pieces demanded	10	11	12	13	14	15
Probability		K+0.02	K+0.05	5K-0.02	7K+0.03	2K+0.02	2K

i. Find the value of K.

ii. Find an expression for her net profit or loss if she manufactures 'm' pieces and demand is 'n' pieces.

Consider separately the two cases -- 'n' lesser than or equal to 'm', and 'n' greater than 'm'.

iii. Find the net profit or loss, assuming that she manufactures 12 pieces.

iv. Find the expected net profit.

v. Calculate expected profit for different levels of production.

Solution

i. The probability of a distribution cannot exceed 1. Using the principle,

(K+0.02) + (K+0.05) + (5K-0.02)+ (7K+0.03) + (2K+0.02) + 2K = 1

18K + 0.1 = 1

K = 0.90/18 = 0.05

Now we can compute the probability distribution of the demand for any day as this:

No. of pieces demanded :	10	11	12	13	14	15
Probability :	0.07	0.10	0.23	0.38	0.12	0.10

ii. If she manufactures 'm' pieces on any day, the cost is Rs.2m. If the number of pieces demanded on any day 'n' is less than or equal to peices produced 'm', then all the pieces demanded are sold, and the sale proceeds is Rs.3n. But, if the number of pieces demanded on any day 'n' is greater than the pieces produced 'm', then the maximum sales is limited to 'm' and thus the sale proceeds is Rs.3m.

Profit = Sales proceeds - Cost

= Rs.3n - Rs.2m, if 'n' is less than or equal to 'm'

= Rs.3m - Rs.2m = Rs. m, if 'n' is greater than 'm'

iii. No. of pieces produced, m = 12. Lets apply the finding in (ii).

n	Profit (Sales proceeds - Cost)
10	10x3 - 12x2 = Rs. 6
11	11x3 - 12x2 = Rs.9
12	12x3 - 12x2 = Rs.12
13	Rs.12
14	Rs.12
15	Rs.12

We notice that for the first three cases (where n = 10, 11, 12), n is less than or equal to m (12). For the remaining three cases (where n = 13, 14, 15), n is greater than m.

iv. Expected net profit E(X) = Profits x Probabilities

= 6 x 0.07 + 9 x 0.10 + 12 x 0.23 + Rs.12 x 0.38 + Rs.12 x 0.12 + Rs.12 x 0.10

= 0.42 + 0.90 + 2.76 + 5.46 + 4.56 + 1.44 + 1.20

= Rs. 16.74

v. Profit for m

Demand (m)	Probability	Production					
(n)		10	11	12	13	14	15
10	0.07	10	8	6	4	2	0
11	0.10	10	11	9	7	5	3
12	0.23	10	11	12	10	8	6
13	0.38	10	11	12	13	11	9
14	0.12	10	11	12	13	14	12
15	0.10	10	11	12	13	14	15
Expected net profit(Rs.)		10.00	10.79	11.28	11.08	9.74	8.04

Expected net profit, when production (m) is 10,

= 10 (0.07 + 0.10 + 0.23 + 0.38 + 0.12 + 0.10)

= Rs. 10.00 x 1

Expected net profit, when production (m) is 11,

= 8 x 0.07 + 11 x (0.10 + 0.23 + 0.38 + 0.12 + 0.10)

= Rs. 10.79

Expected net profit, when production (m) is 12,

= 6 x 0.07 + 9 x 0.10 + 12 x (0.23 + 0.38 + 0.12 +0.10)

= Rs. 11.28

Expected net profit, when production (m) is 13,

= 4 x 0.07 + 7 x 0.10 + 10 x 0.23 +13 x (0.38 + 0.12 + 0.10)

= Rs. 11.08

Expected net profit, when production (m) is 14,

= 2 x 0.07 + 5 x 0.10 + 8 x 0.23 + 11 x 0.38 + 14 x (0.12 + 0.10)

= Rs. 9.74

Expected net profit, when production (m) is 15,

= 0.0 x 0.07 + 3 x 0.10 + 6 x 0.23 + 9 x 0.38 + 12 x 0.12 + 15 x 0.10

= Rs. 8.04

From the expected net profit given in table, we conclude that the maximum expected profit is Rs. 11.28 which happens when production (m) is 12. Hence, the production of 12 pieces per day will optimise Ravali's food stall enterprise's expected profit.

Mathematical Expectation and Variance

The concept, mathematical expectation also called the expected value, occupies an important place in statistical analysis. The expected value of a random variable is the weighted arithmetic mean of the probabilities of the values that the variable can possibly assume. Robert L. Brite has defined the mathematical expectation as: It is the expected value of outcome in the long run. In other words, it is the sum of each particular value within the set (X) multiplied by the probability. Symbolically

The concept of mathematical expectation was originally applied to games of chance and lotteries, but the notion of an expected value has become a common term in everyday parlance. This term is popularly used in business situations which involve the consideration of expected values.

Illustration 31

Mr. Reddy, owner of petrol bunk sells an average of Rs. 80,000 worth of petrol on rainy days and an average of Rs. 100,000 on clear days. Statistics from the Meteorological Department show that the probability is 0.83 for clear weather and 0.17 for rainy weather. Find the expected value of petrol sale and variance.

Solution

We are given

X1 = 80,000

P_1= 0.17

X_2 = 100,000

P_2 = 0.83

The expected value of petrol sale E(X)

= P1 X1 + P_2 X2

= 0.17 x 80000 + 0.83 x 100000 = 136000 + 83000 = Rs. 96600.

= $(80000)^2$ x 0.17 + $(100000)^2$ x 0.83

= 1088000000 + 8300000000 = 9388000000

= 9388000000 - $(96600)^2$ = 9388000000 – 9331560000 = 56440000

S = Sqrt(Variance) = Rs. 7513

Illustration 32

There are three alternative proposals before a businessman to start a new project.

Proposal A - Profit of Rs. 4 lakhs with probability of 0.6 or a loss of Rs. 70,000 with probability of 0.35

Proposal B - Profit of Rs. 8 lakhs with probability of 0.4 or a loss of Rs. 2 lakhs with probability of 0.6

Proposal C - Profit of Rs. 4.5 lakhs with probability of 0.8 or a loss of Rs. 50,000 with probability of 0.2

For maximising profits and minimising loss which proposal he should prefer?

Solution

Calculate the expected value of each proposal.

Formula

E(X) = P1X1 + P_2X2 + ...

Proposal A:

Expected value = Rs. 400,000 x 0.6 - Rs. 70,000 x 0.35 = Rs. 240,000 - 24,500 = Rs. 215,500

Proposal B :

Expected value = Rs.800,000 x 0.4 - Rs. 200,000 x 0.6 = Rs. 320,000 - Rs. 120,000 = Rs. 200,000

Proposal C :

Expected value = Rs. 450,000 x 0.8 - Rs. 50,000 x 0.2 = Rs. 360,000 - Rs. 10,000 = Rs. 350,000

The maximum expected value is Rs. 350,000 which is in case of proposal C. Hence the business man should prefer proposal C.

Illustration 33

The probability that there is at least one error in an accounts statement prepared by A is 0.3, and for B and C the probailities are 0.5 and 0.4 respectively. A, B and C prepared 10, 16 and 20 statements respectively. Find the expected number of correct statements in all and the standard deviation.

Solution

We are given,

$X_1 = 10$; $P_1 = 1 - 0.3 = 0.7$

$X_2 = 16$; $P_2 = 1 - 0.5 = 0.5$

$X_3 = 20$; $P_3 = 1 - 0.4 = 0.6$

$E(X) = P_{1X}1 + P_{2X2} + P_{3X3}$

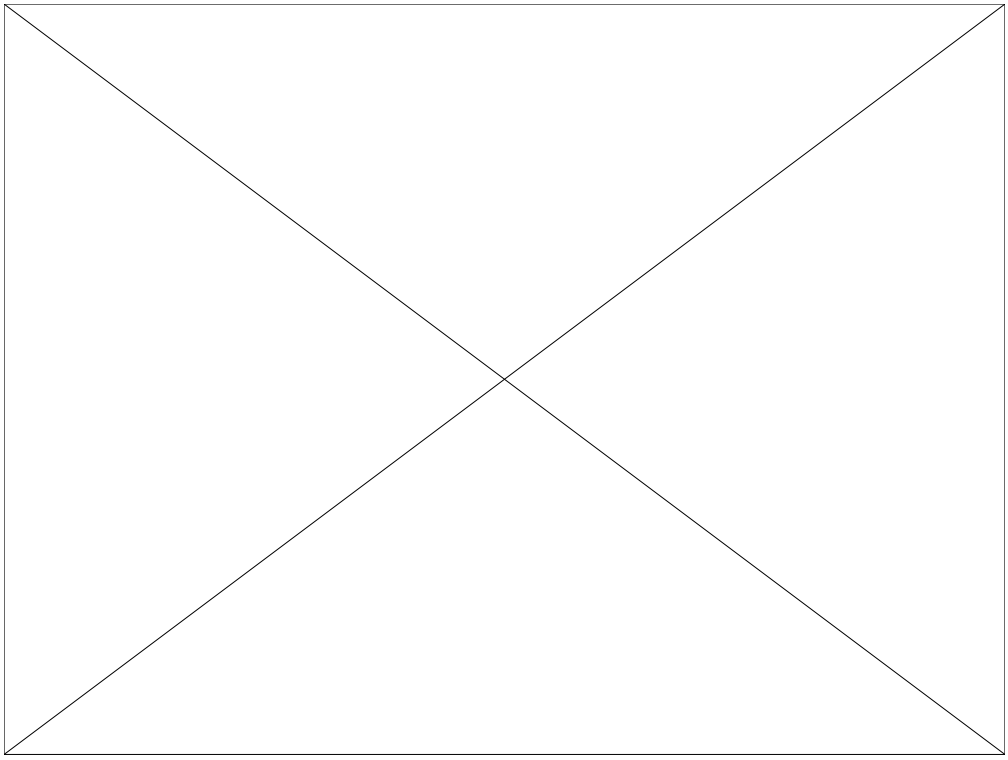
$= 0.7 \times 10 + 0.5 \times 16 - 0.6 \times 20 = 7 + 8 + 12 = 27$

Thus, expected number of correct statements are 27.

$= (10)^2 \times 0.7 + (16)^2 \times 0.5 + (20)^2 \times 0.6 = 70 + 128 + 240 = 438$

$= 438 - (27)^2 = 438 - 729 = -291$

SD = Sqrt(Variance) = $-\sqrt{291} = -17.06$



Please use headphones

- End of Chapter -

LESSON - 7

CONCEPT AND USES OF DISTRIBUTION

Inferences about the characteristics of population can be drawn through (a) observed or experimental frequency distributions and (b) theoretical or probability distributions. In frequency distribution, measures like average, dispersion, correlation, etc. are studied based on the actual data so that it is possible to deduce inferences which indicate both (i) the nature and form of the sample data and (ii) help in formulating certain ideas about the characteristics of population. In population, the values of variable may be distributed according to some definite probability law, and the corresponding probability distribution is known as Theoretical Probability Distribution. Such probability distributions are based on prior considerations or previous experience and are more scientific approach to draw inferences about the characteristics of population fitting a mathematical model or a function of the form Y = P(X) to the given data.

We have defined the mathematical expectation, random variable, and probability distribution function and also discussed these. In the present section, we will cover the following univariate probability distributions:

(i) Binomial Distribution,

(ii) Poisson Distribution, and

(iii) Normal Distribution.

The first two distributions are discrete probability distributions and the third one is a continuous probability distribution.

Binomial Distribution

Binomial distribution is named after the Swiss mathematician James Bernoulli who innovated it. The binomial distribution is used to determine the probability of success or failure of the one set in which there are only two equally likely and mutually exclusive outcomes. This distribution can be used under specific set of assumptions:

i. The random experiment is performed under the finite and fixed number of trials.

ii. The outcome of each trail results in success or failure.

iii. All the trails are independent in the sense the outcome of any trail is not affected by the preceding or succeeding trials.

iv. The probability of success or failure remains constant from trial to trial.

The success of an event is denoted by 'p' and its failure by 'q'. Since the binomial distribution is a set of dichotomous alternatives i.e. successes or failures and thus, p + q = 1. Hence, (q + p) are the terms of binomial. By expanding the binomial terms, we obtain probability distribution which called the binomial probability distribution or simply the binomial distribution. It is defined as

$$P(r) = P(X=r) = {}^{n}C_r \, q^{n-r} p^r$$

By expanding the terms of binomial $(q+p)^n$, we get

$$(q+p)^n = {}^{n}C_0 q^n p^0 + {}^{n}C_1 q^{n-1} p^1 + {}^{n}C_2 q^{n-2} p^2 + \ldots {}^{n}C_n q^0 p^n$$

or

$$(q+p)^n = q^n + {}^{n}C_1 \, q^{n-1} p + {}^{n}C_2 \, q^{n-2} p^2 + \ldots p^n$$

Where ${}^{n}C_0$, ${}^{n}C_1$, ${}^{n}C_2$ ...are called 'binomial coefficients'.

For example, the binomial expansion of $(q + p)^5 = q^5 + 5\, q^4\, p + 10\, q^3\, p^2 + 10\, q^2\, p^3 + 5\, q\, p^4 + p^5$

and 1, 5, 10, 10, 5, 1 are coefficients of the binomial.

Rules of binomial expansion

In binomial expansion, the rules should be noted.

(i) The number of terms in a binomial expansion is n + 1

(ii) The binomial coefficients for n + 1 terms of the distribution are symmetrical, ascending up to the middle of the series, and then descending. When n is odd number, n + 1 becomes even and the binomial coefficients of the two central terms are identical.

The constants of binomial distribution are:

Mean of binomial distribution = np

S.D. of binomial distribution = $\sqrt{npq}$

First moment of binomial distribution = $\mu_1 = 0$

Second moment of binomial distribution = $\mu_2 = npq$

Third moment of binomial distribution = $\mu_3 = npq(q-p)$

Fourth moment of binomial distribution = $\mu_4 = 3n^2p^2q^2 + npq(1-6pq)$

$\beta_1 = (q - p)^2 / npq$

$\beta_2 = 3 + (1 - 6pq) / npq$

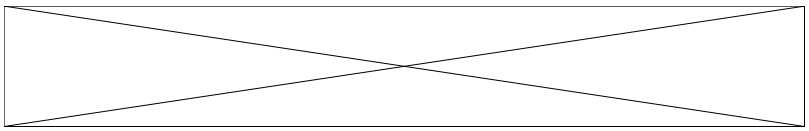
Procedure for Binomial Distribution

Step 1. Determine the values of p and q. If one of these values is known, the other can be found by their relationship -

p = (1 - q) and q = (1 - p)

Step 2. Expand the binomial $(q + p)^n$

Step 3. Multiply each term of the expanded binomial by N (total frequency) in order to obtain the expected frequency in each category.



Please use headphones

Illustration 34

A coin is tossed six times. What is the probability of obtaining four or more heads.

Solution

In a toss of an unbiased coin, the probability of head as well as tail is equal, i.e, $p = q = 1/2$

By expanding the terms $(q + p)^6$, we get various possibilities for all the events.

$(q + p)^6 = q^6 + 6q^5p + 15q^4p^2 + 20q^3p^3 + 15q^2p^4 + 6qp^5 + p^6$

Required probability of obtaining four heads is

$15q^2p^4 = 15 \times (1/2)^2 \times (1/2)^4 = 15/64$

Required probability of obtaining five heads is

$6q^1p^5 = 6 \times (1/2)^1 \times (1/2)^5 = 6/64$

Required probability of obtaining six heads is

$p^6 = (1/2)^6 = 1/64$

Hence required probability of obtaining four or more heads is

Illustration 35

Eight coins are thrown simultaneously. Show that the probability of obtaining at least 6 heads is 37/256.

Solution

Probability of getting head and tail are denoted by p and q respectively. In case of unbiased coin, $p = q = 1/2$.

The probability of r successes i.e., getting heads in 8 trials is given by

$p(r) = {}^nC_r\, q^{n-r}\, p^r$

$p(r) = {}^8C_r\, q^{8-r}\, p^r = {}^8C_r\, (1/2)^{8-r}\, (1/2)^r = {}^8C_r\, (1/2)^8 = {}^8C_r\, /256$

Probability of obtaining at least 6 heads = Probability of obtaining 6 heads or 7 heads or 8 heads

$= {}^8C_6\, /256 + {}^8C_7\, /256 + {}^8C_8\, /256 = 1/256\, ({}^8C_6 + {}^8C_7 + {}^8C_8) = 1/256(28+8+1) = 37/256$

Illustration 36

The incidence of occupational disease in an industry is such that the workman have a 20% chance of suffering from it. What is the probability that out of six workmen, three or more will contact the disease?

Solution

Probability of a man suffering from disease is p = 20% = 1/5

q = 1 - p = 1 - 1/5 = 4/5

Probability of 3 or more men suffering from disease is termed in binomial expansion of $(1/5 + 4/5)^6$

Probability of men contracting the disease

$= q^6 + {}^6C_1\, q^5 p + {}^6C_2\, q^4 p^2 + {}^6C_3\, q^3 p^3 + {}^6C_2\, q^2 p^4 + {}^6C_1\, qp^5 + p^6$

And, probability of 3 or more men contracting the disease

$= \cancel{q^6 + {}^6C_1\, q^5 p + {}^6C_2\, q^4 p^2 +} {}^6C_3\, q^3 p^3 + {}^6C_2\, q^2 p^4 + {}^6C_1\, qp^5 + p^6$

$= 20q^3p^3 + 15q^2p^4 + 6qp^5 + p^6$

$= 20\,(4/5)^3\,(1/5)^3 + 15\,(4/5)^2\,(1/5)^4 + 6\,(4/5)\,(1/5)^5 + (1/5)^6$

= 20 x 64/15625 + 15 x 16/15625 + 6 x 4/15625 + 1/15625

= 1545/15625 = 0.09888

Illustration 37

a. For a binomial distribution, mean = 7 and variance =11. Comment.

b. If the probability of a defective item among 600 items is 1/12, find

i) the mean

ii) the variance

Solution

a. We know that Mean = np, and Variance = npq

np = 7 and npq = 11

(7)q = 11

q = 11/7

b. Probability of defective item (p), is given as 1/12 and n = 600

q = 1 - p = 1 - 1/12 = 11/12

By using binomial law,

i) Mean = np = 600 x 1/12 = 50

ii) Variance = npq = 50 x 11/12 = 45.83

Illustration 38

Nine coins are tossed one at at a time, 512 times. Number of heads observed is recorded at each throw, and the results are given below. Find the expected frequencies. What are the theoretical values of mean and standard deviation? Also calculate the mean and standard deviation of the observed frequencies.

No. of heads at a throw	Frequency
0	4
1	10
2	45
3	115
4	139
5	105
6	65
7	19
8	8
9	2

Solution

Probability of getting a head in a single throw, p = 1/2

Therefore, probability of not getting a head in a single throw, q = 1 - p = 1 - 1/2

Also, we are given n = 9, and N = 512

The expected frequencies can be calculated by expanding $512\ (1/2 + 1/2)^9$

No. of heads	Expected frequencies = $N\ {}^nC_r\ q^{n-r}\ p^r$
0	$512 \times (1/2)^9$ = 1

1	$512 \times {}^9C_1 (1/2)^8 (1/2)^1$	= 9
2	$512 \times {}^9C_2 (1/2)^7 (1/2)^2$	= 36
3	$512 \times {}^9C_3 (1/2)^6 (1/2)^3$	= 84
4	$512 \times {}^9C_4 (1/2)^5 (1/2)^4$	= 126
5	$512 \times {}^9C_5 (1/2)^4 (1/2)^5$	= 126
6	$512 \times {}^9C_6 (1/2)^3 (1/2)^6$	= 84
7	$512 \times {}^9C_7 (1/2)^2 (1/2)^7$	= 36
8	$512 \times {}^9C_8 (1/2) (1/2)^8$	= 9
9	$512 \times {}^9C_9 (1/2)^9$	= 1
Total		512

Mean value of expected probability distribution = np = 9 x 1/2 = 4.5

Standard Deviation

Now, we calculate the mean and standard deviation of observed or actual frequency distribution.

Arrange data:

X	dx	Frequency	F.dx	$F.dx^2$
0	-5	4	- 20	100
1	-4	10	-40	160
2	-3	45	-135	4d5
3	-2	115	-230	460
4	- 1	139	- 139	139
5	0	105	0	0
6	+ 1	65	+ 65	65
7	+ 2	19	+ 38	76
8	+ 3	8	+ 24	72
9	+ 4	2	+8	32
	-5	512	- 429	1509

$$\text{Mean} = A + \frac{\sum Fdx}{N} = 5 - \frac{429}{512} = 5 - 0.8379 = 4.1621$$

$$\text{S.D} = \sqrt{\frac{\sum Fdx^2}{N} - \left(\frac{\sum Fdx}{N}\right)^2} = \sqrt{\frac{1509}{512} - \left(\frac{429}{512}\right)^2}$$

$$= \sqrt{2.9473 - 0.7021} = 1.4984$$

Remark

For frequency distribution, the mean and standard deviation are 4.1621 and 1.4984 respectively while the figures for probability frequency distribution are 4.50 and 1.50. The probability frequency is more scientific and mathematical model so that the arriving results are more accurate and precise. For example, if we substitute the mean of observed frequency distribution, we get

Mean = np = 4.1621

So, p = 4.1621 / 9 = 0.462455

Similarly Variance = npq = SD^2 = $(1.4984)^2$ = 2.2452

So, q = 2.2452 / 4.1621 = 0.539439

The value of p + q = 1. But in our example, the value of (p + q) calculated based on the observed frequency distribution is 0.462455 + 0.539439) = 1.001894

Hence, the binomial probability distribution is more scientific and describes the real life of an event.

Illustration 39

Given data shows the number of seeds germinating out of 10 on damp filter for 100 sets of seeds. Fit a binomial distribution

X	=	0	1	2	3	4	5	6	7	8	9	10
Y	=	6	20	30	16	12	8	4	3	1	0	0

Solution

First, find out the terms of binomial expansion i.e. p and q

X	F	Fx
0	6	0
1	20	20
2	30	60
3	16	48
4	12	48
5	8	40'
6	4	24
7	3	21
8	1	8
9	0	0
10	0	0
	100	269

Mean = np

$p = np/n = 2.69 / 10 = 0.269$

$q = 1 - p = 1 - 0.269 = 0.731$

Hence, Binomial Distribution = $100 (0.731 + 0.269)$. By expanding $100(0.731 + 0.269)^{10}$, we get the expected frequencies for 0, 1, 2, ...10 and the results are tabulated below:

X	Expected Frequencies $N \times {}^{n}C_r\, q^{n-r}\, p^r$	
0	$100 \times (0.731)^{10}$	= 4.36
1	$100 \times {}^{10}C_1 (0.731)^9 (0.269)$	= 16.03
2	$100 \times {}^{10}C_2 (0.731)^8 (0.269)^2$	= 26.54
3	$100 \times {}^{10}C_3 (0.731)^7 (0.269)^3$	= 26.05
4	$100 \times {}^{10}C_4 (0.731)^6 (0.269)^4$	= 16.80
5	$100 \times {}^{10}C_5 (0.731)^5 (0.269)^5$	= 7.40
6	$100 \times {}^{10}C_4 (0.731)^4 (0.269)^6$	= 2.27
7	$100 \times {}^{10}C_3 (0.731)^3 (0.269)^7$	= 0.48
8	$100 \times {}^{10}C_2 (0.731)^2 (0.269)^8$	= 0.07
9	$100 \times {}^{10}C_1 (0.731)^1 (0.269)^9$	= 0
10	$100 \times {}^{10}C_0 (0.269)^{10}$	= 0
		100.00

- End of Chapter -

LESSON - 8

POISSON DISTRIBUTION

Poisson Distribution was found by French mathematician Simeon D. Poisson. This distribution describes the behaviour of rare events and has been known as the Law of Improbable events. Poisson distribution is a discrete probability distribution and is very popularly used in statistical inferences. The binomial distribution can be used when only the sample space (number of trials n) is known, while the Poisson distribution can study when we know the mean value of occurrences of an event without knowing the sample space. Mathematically, the Poisson distribution is in limiting form of binomial distribution as n (number of trials) tends to infinity and p (success) approaches zero, in that way np= m remains constant. Such distribution is

fairly common. Under the conditions, n is infinity, p approaches zero and np = m remains constant, the Binomial distribution function tends to Poisson probability function which is given below (definition of Poisson probability distribution).

where x is the number of succeses (occurrences of an event), m = np and e = 2.71823 (the base of natural logarithm). m is called the parameter of the Poisson ditsribution. The standard deviation is $\sqrt{m}$.

Application and Uses

Poisson distribution can explain the behaviour of the discrete 'random variables where the probability of occurrence of events is very small and the number of trials is sufficiently large As such, this distribution has found application in many fields like Queuing theory, Insurance, Biology, Physics, Business, Economics, Industry etc. The practical areas where the Poisson distribution can be used is listed below. It is used in...

1. waiting-time problems to count the number of incoming telephone calls or incoming customers to market or number of traffic arrivals such as truck at terminals, aeroplanes at airports, ships at docks,

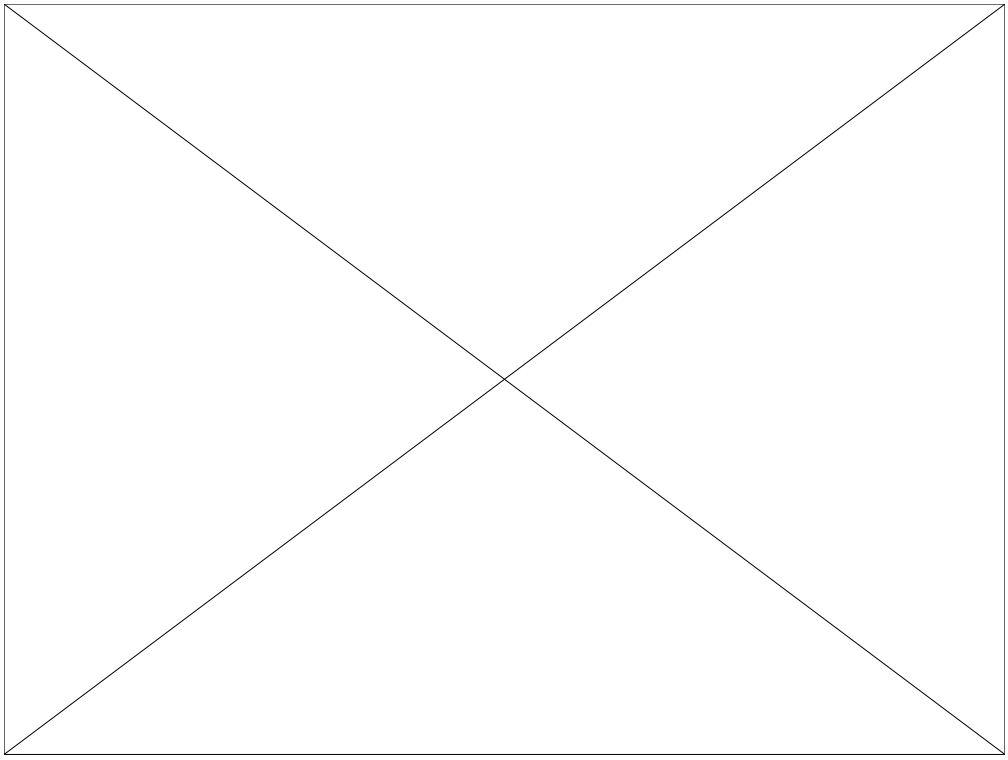
2. insurance problems to count the number of casualties,

3. Biology to count the number of bacteria,

4. Physics to count the number of disintegrating of a radioactive element per unit of time,

5. business to count defects per unit of production.

In addition to the above, the Poisson distribution can also use in things like counting number of accidents taking place per day, in counting number of suicides in a particular day, or persons dying due to a rare disease such as heart attack or cancer or snake bite or plague, in counting number of typographical errors per page in a typed or printed material etc.



Please use headphones

Illustration 40

An average number of phone calls per minute into the switch-board of Reddy Company Limited between the hours of 10 AM to 1 PM is 2.5. Find the probability that during one particular minute there will be (i) no phone calls at all, (ii) exactly 3 calls and (iii) at least 5 calls.

Solution

Let us denote the number of telephone calls per minute by X. Then X follows Poisson distribution with mean distribution, m = 2.5. The Poisson probability function is:

(i) Probability of no calls =

$p(x=0) = e^{-2.5} (2.5)^0 / 0! = e^{-2.0} \times e^{-0.5} \times 1 / 1 = 0.13534 \times 0.6065 = 0.08208$

Note:

Refer Table for $e^{-2.5}$. We cannot get table value for 2.5. First we have to find the value for $e^{-2.0}$ (= 0.13534) and then for $e^{-0.5}$ (= 0.6065). Now, multiply them to get 0.08208).

(ii) Probability of exactly 3 calls is

(iii) Probability of atleast 5 calls is

Illustration 41

The mistakes per page were observed in a book, fit a Poisson distribution. The data are:

No. of mistakes per page (x)	0	1	2	3	4
No. of times the mistakes occurred (F)	211	90	19	5	0

Solution

First, calculate the value of mean distribution

$m = \sum Fx / N$

$= [(0 \times 211) + (1 \times 90) + (2 \times 19) + (3 \times 5) + (4 \times 0)] / 325 = 0.44$

Apply Poisson distribution,

$p(x=r) = N \times (e^{-m} m^r)/r!$

$e^{-0.44} = 0.644$

Expected frequencies:

No. of mistakes per page (r)	Expected frequencies	
0	$325 \times 0.644 (0.44)^0$	= 209.34
1	$325 \times 0.644 (0.44)^1$	= 92.09
2	$325 \times 0.644 (0.44)^2/2!$	= 20.26
3	$325 \times 0.644 (0.44)^3/3!$	= 2.98
4	$325 \times 0.644 (0.44)^4/4!$	= 0.33
		325.00

Illustration 42

The components processed by a machine have been found to have some defects. 40 components were selected at random and the number of defects in each of them were noted. The data is:

4	1	2	2	1	3	2	4
0	1	3	2	4	3	0	1
2	3	0	1	1	0	2	2
4	0	2	1	5	1	1	3
0	2	5	2	3	0	1	2

(a) Determine the probability distribution of the random variable, number of defects in a component and frequency distribution based on the Poisson distribution.

(b) Verify whether a Poisson distribution can be assumed Chi-square test.

Solution

First determine observed frequency.

No. of defects (x)	Tallies	F	Fx
0	𝍸 ||	7	0
1	𝍸 𝍸	10	10
2	𝍸 𝍸 |	11	22
3	𝍸 |	6	18
4	𝍸	4	16
5	||	2	10
		40	76

$m = \sum Fx / N = 76/40 = 1.9$

Expected frequency of defects containing 'r' defects according to binomial law is given by...

$NP(r) = 40 \times e^{-m} m^r / r!$

$40 \times e^{-m} = 40 \times e^{-1.9} = 40 \times 0.14958 = 5.9832$

Defects (r)	Expected frequency (rounded)	F

0	$5.9833 \times (1.9)^0$	6
1	$5.9833 \times (1.9)^1$	12
2	$5.9833 \times (1.9)^2/2!$	11
3	$5.9833 \times (1.9)^3/3!$	7
4	$5.9833 \times (1.9)^4/4!$	3
5	$5.9833 \times (1.9)^5/5!$	1
		40

Chi Square

$$= (7-6)^2/6 + (10-12)^2/12 + (11-11)^2/11 + (6-7)^2/7 + (4-3)^2/3 + (2-1)^2/1$$

$$= 1/6 + 1/3 + 0 + 1/7 + 1/3 + 1$$

$$= 0.17 + 0.33 + 0.14 + 0.33 + 1 = 1.97$$

For 4 d.f at 0.05 = 9.49

The calculated value is less than the table value, hence Poisson distribution provides a good fit to the data.

Note:

Without referring to Table, we can calculate the value of e. For example, $e^{-0.2} = 0.8188$.

This can be computed as:

$$-0.2 \log_e$$

$$= -0.2 \log_{10} 2.7183$$

$$= -0.08686$$

Antilog [-0.08686] = Antilog -1.91314 = 0.8187

Normal Distribution

The Binomial and Poisson distributions discussed in preceeding paragraphs are most useful theoretical distribution for discrete variables i.e. occurrence of disjoint events. A more suitable distribution for dealing with the variable whose magnitude is continuous is normal distribution. It is also called the normal probability distribution.

Uses

1. It aids solving many business and economic problems including the problems in social and physical sciences. Hence, it is cornerstone of modern statistics.

2. It becomes a basis to know how far away and in what direction a variable is from its population mean.

3. It is symmetrical. Hence mean, median and mode are identical and can be known.

4. It has only one maximum point at the mean, and hence it is unimodel (i.e. only one mode).

Definition

In mathematical form, the normal probability distribution is defined by:

Where,

Y^{ε} = ordinate of the curve at a point x

N = number of items in the frequency distribution

θ = standard deviation of the distribution

π = 22/7 = 3.14159

e = 2.71828

The above equation can also be written as $Y^{\varepsilon} = Y^{o\varepsilon}(-x^2/2\boldsymbol{\theta}^2)$

where

$Y^o = N/\theta\ \sqrt{2\pi}$

If the normal curve is in terms of standard deviation units, it is called normal deviate. The normal deviate at the mean will be zero viz. $x = x/\theta = 0/\theta = 0$, where $x\text{-}\bar{x}$ at values equal to 10, 20, and 30 will be respectively 1θ, 2θ and 3θ (to the left of the mean, these units will be negative as x variates will be less than the value of mean, and to the right of the mean, these units will be positive as x variates will be greater than the value of mean). This is known as changing to standardized scale. In equation the changing to standardized scale is written as

The normal curve is distributed as under:

a. Mean ± 1θ covers 68.27% area; 34.135% on either side of the mean

b. Mean ± 2θ covers 94.45% area; 47.725% on either side of the mean

c. Mean ± 3θ covers 99.73% area; 49.865% on either side of the mean

Method of Ordinates

To make a curve on graph, we need the frequencies and the values of variable which represent on the ordinate (Y-axis) and abscissa (X-axis) respectively. Hence, in order to fit a curve we must know the ordinates (i.e. frequencies) at the various points of the abscissa scale. Find the $\bar{x}$, N and class interval, if any, of the observed distribution. Then calculate Y^o

$Y^o = Ni / \sqrt{2\pi} = Ni/2.5071$

$$Y^o = 0.399 (Ni/\theta)$$

This gives mean ordinate.

Illustration 41

The customer accounts at the Departmental Store have an average balance of Rs. 120 and standard deviation of Rs. 40. Assuming that the account balances are normally distributed, find

i. What proportion of the accounts is over Rs. 150.

ii. What proportion of the accounts is between Rs. 100 and Rs. 150.

iii. What proportion of the accounts is between Rs. 60 and Rs. 90.

Solution

We are given

$\bar{x} = 120$, and $\theta = 40$

Formula

i. Proportion of accounts over Rs.150 (x = 150)

So, $Z_{150} = (150-120) /40 = 0.75 = 75\%$

Referring to Z-table the area under, Z = 0.2734 (See table for 0.75).

We have to find probability of items falling to the right of Z = 0.74 i.e. over Rs. 150. Deduct the value of 0.2734 from the total probability to the right origin, 0.5 - 0.2734 = 0.2266. Hence, 22.66 per cent of the accounts have a balance in excess of Rs. 150.

ii. Proportion of the accounts between Rs. 100 and Rs. 150

Z_{100} = (100-120) /40 = (-)0.50

Z_{150} = (150-120) /40 = 0.75

Referring to Z-table,

Area between Z_x = (-)0.50 and Z_x = 0.75

= 0.1915 + 0.2734 = 0.4649

Therefore 46.49 percent of the accounts have an average between Rs. 100 and Rs.150

iii. Proportion of the accounts between Rs. 60 and Rs. 90

Z_{60} = (60-120) /40 = (-)1.5

Z_{90} = (90-120) /40 = (-)0.75

Referring to Z-table

Area corresponding to Z = (-)1.50 = 0.4332

Area corresponding to Z = (-) 0.75 = 0.2734

Area between Z_{60} = (-)1.50 and Z_{90} = (-)0.75 is 0.4332 - 0.2734 = 0.1598

Thus, 15.98 per cent of the accounts have an average between Rs. 60 and Rs. 90.

Illustration 42

In a public examination 6000 students have appeared for statistics. The average mark of them was 62 and standard deviation was 10. Assuming the distribution is normal, obtain the number of students who might have obtained (i) 80 percent or more, (ii) First class (i.e. 60 per cent), (iii) secured less than 40 per cent and (iv) if there are only 150 vacancies, find the minimum mark that one should secure to get selected against a vacancy.

Solution

We are given

$\bar{x}$ = 62 and θ = 10

i. Z_{80} = (80-60) /10 = 2.0

Referring to Z-table for Z = 2.0 gives 0.4773

Number of students secure 80 per cent or more is 0.5 - 0.4773 = 0.0227, i.e. 2.27 per cent of students. Thus, 600 x 0.0227 = 136 students secured 80 per cent or more.

ii. Z_{60} = (60-62) /10 = (-)0.2

For Z = (-)0.2, value = 0.0783

Hence, 0.0793 x 6000 = 476 students scored more than 60%

iii. Z_{40} = (40-62) /10 = (-)2.2

For Z = (-)2.2, value = 0.4861

Hence, no. of students who scored less than 40% marks is 6000 x (0.5 - 0.4861) = 6000 x 0.139 = 8

iv. No. of vacancies = 150

Area under which top 150 students fall is:

150/6000 = 0.025

Area under which the students who secure more than 150 rank is 0.5 - 0.025 = 0.475. In other words, the students who secure more than 150 ranks fall under the area of 0.475. The Z value corresponding to 0.475 of Z-tab;e is 1.96 (see Table)

Z = (X - 62) / 10 = 1.96

X = 19.6 + 62 = 81.6

The minimum mark that one should obtain to get selected against a vacancy is 81.6

EXERCISES

1. What do you mean by probability. Discuss the importance of probability in statistics?

2. What is meant by mathematical expectation? Explain it with the help of an example?

3. What is Bayes' theorem? Explain it with suitable example?

4. What is meant by the Poisson distribution? What are its uses?

5. Explain the terms

i. Mutually exclusive events

ii. Independent and dependent events

iii. Simple and compound events

iv. Random variable

v. Permutation and combination

vi. Trial and event

vii. Sample space

6. 3 balls from an urn containing 6 white and 4 black balls are drawn. Find the probability that 2 are white and 1 is black...(i) if each ball is returned before the next is drawn, (ii) if the three balls are drawn successively without replacement. (Ans: i. 1/2; ii. 1/6)

7. A bag containing 8 white, 6 red and 4 black balls. Three balls are drawn at random. Find the probability that they willbe white. (Ans: 56/816)

8. A bag contains 4 white and 8 red balls, and a second bag 3 white and 5 black balls. One of the bags is chosen at random and a draw of 2 balls is made it. Find the probability that one is white and the other is black. (Ans: 0.510)

9. A class consists of 100 students, 25 of them are girls and the remaining are boys, 35 of them are rich and 65 poor, 20 of them are fair glamor What is the probability of selecting a fair glamor rich girl. (Ans: 0.0175)

10. Three persons A, B and C are being considered for the appointment as Vice-Chancellor of a University whose chances of being selected for the post are in the proportion 5:3:2 respectively. The probability that A if selected will introduce democratisation in the University strut is 0.3, the corresponding probabilities for B and C doing the same are respectively 0.5 and 0.7. What is the probability that democratisation would be introduced in the University. (Ans: 0.44)

11. The probability that a trainee will remain with a company is 0.65. The probability that an employee earns more than Rs. 1800 per year is 0.60. The probability that an employee is a trainee who remained with the company or who earns more than Rs.18000 per year is 0.75. What is the probability than an employee earns more that Rs. 18000 per year given that he is a trainee who stayed with the company.

12. In a bolt factory, machines A, B, C produce 30 per cent, 40 per cent and 30 per cent respectively. Of their output 3, 4, 2 per cents are defective bolts. A bolt is drawn

at random from the product and is found to be defective. What are the probabilities that it was produced by machines A, B and C. (Ans: 0.29, 0.52, 0.19)

13. A factory produces a certain type of output by two types of machines. The daily production are: Machine I - 4000 units and Machine II - 5500 units. Past records show that defectives for the output produced by Machine I and Machine II are 1.4 per cent and 1.9 per cent respectively. An item is drawn at from the day's production and is found to be defective. What is the probability that it comes from the output of (i) Machine I and (ii) Machine II. (Ans: (i) 0.3489 (ii) 0.6511)

14. Dayal company estimates the net profit on a new product it is launching to be Rs. 5 lakhs, during the first year if it is successful, Rs. 3 lakhs if it is made moderately successful. The company assigning the probabilities to the first year prospects for the product are: Successful - 0.18, moderately successful -0.22. What are the expected profit and standard deviation of first year net profit for his product. (Ans: Profit 0.66 lakhs; S.D. Rs. 2.719 lakhs)

15. A systematic sample of 100 passes was taken from the concise Oxford Dictionary and the observed frequency distribution of foreign words per page was found to be as follows :

No. of foreign words per page (x) :	0	1	2	3	4	5	6
Frequency:	48	27	12	7	4	1	1

Calculate the expected frequencies using Poisson distribution. Also calculate the variance of fitted distribution. (Ans: 37, 37, 18, 16, 2, 0, 0; Variance = 0.99)

16. Income of a group of 10000 persons were found to be normally distributed with mean Rs. 750 per month and standard deviation Rs. 50. Of third group, about 95 per cent had income exceeding Rs. 668 and only 5 per cent had income exceeding Rs. 832. What was the lowest income among the richest 100. (Ans: Rs. 866)

REFERENCE BOOKS

Agarwal, B.L. 'Basic *Statistics',* Wiley Eastern Ltd , New Delhi, 1994.

Chance, W., *'Statistical Methods for Decision - making',* Irwin Inc., Homewood, 1969.

Gopikuttam, G., *'Quantitative Methods and Operations Research', Himalaya Publishing House,* Bombay.

Gupta, S.P., *'Statistical Methods',* Sultan Chand and Co., New Delhi.

Levin, R.. *'Statistics for Management',* Prentice - Hall of India, New Delhi, 1984.

Monga, G.S.,*'Mathematics and Statistics for Economies'* Vikas Publishing House, New Delhi, 19

Reddy, C.R., *'Quantitative Methods for Management Decision',* Himalaya Publishing House, Bomay, 1990.

-End of Chapter -

LESSON - 9

STATISTICAL INFERENCES

Introduction

Human welfare including daily actions of human beings leans heavily on statistics. Inferring valid conclusions for making decision needs the study of statistics and application of statistical methods almost in every field of human activity. Statistics, therefore, is regarded as the science of decision making. The statisticians can commonly categorise the techniques of statistics which are of so diverse into (a) descriptive statistics and (b) inferential statistics (or inductive statistics). The former describes the characteristics of numerical data while the latter describes the judgment based on the statistical analysis. In other words, the former is process of analysis. In other words, the former is process of analysis whereas the latter is that of scientific device of inferring conclusions. Both are the systematic methods of drawing satisfactory valid conclusions about the totality (i.e. population) on the basis of examining a part of population, termed as sample. The process of studying the sample and then generalising the results to the population needs a scientific investigation searching for truth.

Population and Sample

The word population is technical term in statistics, not necessarily referring to people. It is totality of objects under consideration. In other words, it refers to a number of objects or items which are to be selected for investigation. This term as sometimes called the universe. Figure 1. shows the concept of population and a sample in the form of the Venn diagram where population is shown as the universal set and a sample is shown as a true subject of the population.

A population containing a finite number of objects say the students in a college, is called finite population. A population having an infinite number of objects say, heights or weights or ages of people in the country, stars in the sky etc. is known as an infinite population. Having concrete objects say, the number of books in a library, the number of buses or scooters in a district, etc. is called existent population. If the population consists of imaginary objects say, throw of die or coin in infinite numbers of times is referred to hypothetical population.

For social scientist, it is often difficult, in fact impossible to collect information from all the objects or units of a population. He, therefore, interested to get sample data. Selection of a few objects or units forming true representative of the population is termed as sampling and the objects or units selected are termed as sample. On the analysis being derived from the sample data, he generalises to the entire population from which the sample is drawn. The sampling has two objectives which are: (a) obtaining optimum information and (b) getting the best possible estimates of (population) parameters.

Parameter and Statistics

The statistical constants of the population such as population size (N), population mean(m) , population variance (θ^2), population correlation coefficient (p), etc are called parameters. In other words, the values that are derived using population data are known as parameters. Similarly, the values that are derived using sample data are termed as statistics not to be confused with the word statistics meaning data or the science of statistics. The examples for statistics are sample mean *(x)*, sample variance (S^2), sample correlation coefficient (r), sample size (n), etc Obviously 1 statistics are quotients of the sample data whereas parameters are function of the 1 population data. In brief the population constant is called parameter while the 1 sample constant is known as statistics.

Random Sample

Sampling refers to the method of selecting a sub-set of the population for investigation. Selection of objects or units in such a way that each and every object or unit in the population has the chance of being selected is called random sampling. The number of objects or units in the sample is termed as sample size. This size should neither be too big nor too small but should be optimum. Over the census method, the sample method has distinct merits, which R.A. Fisher sums up thus: Speed, economy, adaptability and scientific. The right type of sampling plan is of paramount importance in execution of a sample survey in accordance with the objectives and scope of investigation the sampling techniques are broadly classified into (a) random sample, (b) non-random sample and (c) mixed sample.

The term random (or probability) is very widely applicable technique in selecting a sample from the population. All the objects or units in the universe will have an equal chance of being included in the random sample. In other words, every unit or object is as likely to be considered as any other. In this, the process is random in character and is usually representative. Selecting 'n' units out of N in such a way that every one of Ncn samples has an equal chance of being selected. This is done in the ways: (a) random sampling with replacement (rswr) and (b) random sampling without replacement (rswor). The former does permit replacing while the latter does not.

Let x stands for the lift (in hours) of television produced by Konark Company under essentially identical conditions (with the same set of workers working on the same machine using the same type of materials and the same technique). If x1, X2, X3, ...x_n are the lives of n such television, then x1, x2, X3.....x_n may be regarded as a random sample from the distribution of x. The number n is called the size of random sample.

A random sample may be selected either by drawing the chits or by the use of random numbers. The former is a random method but is subject to biases (as can be identified chits). The latter is the best as numbers are drawn randomly.

For example where a population consists of 15 units and a sample of size 6 is to be selected thus since 15 is a two-digit figure, units are numbered as 00, 01, 02, 03 ... 99. Six random numbers are obtained from a two digit random number table They are - 69,36, 75, 91,44 and 86. On dividing 69 by 15, the remainder is 9, hence select the unit on serial number 9. Likewise divide 36, 75, 91 and 44 and 86 by 15. The respective remainders are 6,0,1, 14 and 11. Hence select units of serial numbers 09, 06, 00, 01, 14 and 11. These selected units form the sample.

Sampling Distribution

A function of the random variables x1, x2, x3.....x_n, a statistic is itself a random variable. Hence, a random variable has probability distribution. This probability distribution of a statistic is known as the sampling distribution of the statistics. This distribution describes t he way that a statistic is the function of the random variables. In Practice the sampling distributions which commonly used are the sample mean and the sample variance. These will give a fillip to a number of test statistics for hypothesis testing.

(a) The Sample Mean:

Suppose in a simple random sample of size n picked up from a population, then the sample mean represented by x is defined as

$$\bar{x} = \frac{x1+x2+x3.....+x_n}{n}$$

The sample may be selected with replacement or without replacement. In the former, a number occurring more than once is accepted. A unit is repeated as many times as a random number occurs. the latter, a random number is omitted at an; subsequent stage. The sampling with replacement and without replacement are referred to infinite population and finite population respectively. The expection value of sample mean is the same as the population mean. Thus:

$$E(\bar{x}) = \mu$$

$$E(\bar{x}) = \frac{\sum x}{n}$$

$$E(\mu) = \frac{\sum x}{N}$$

a. The Sample Variance :

Suppose, the simple random sample of size n chosen from a population the sample variance is used to estimate the population variance. In an equation form.

$$\theta^2 = \frac{\left[\sum (x-\mu)^2\right]}{N} \quad \text{rswr}$$

$$S^2 = \frac{\left[\sum x-(\bar{x})\right]}{n} \quad \text{rswr}$$

$$\theta^2 = \frac{\left[\sum (x-\mu)^2 \ldots\ldots\right]}{N} \quad \text{rswor}$$

$$S^2 = \frac{N-n}{N-1} \times \frac{\theta^2}{n} \quad \text{rswor}$$

Standard Error

The standard deviation measures variability variable. The standard deviation of a sampling distribution is referred to standard error (S.E.). It measures only sampling variability which occurs due to chance or random *forces,* in estimating a population parameter. The word error is used in place of deviation to emphasize that the variation among sample statistic is due to sampling errors. If 0 is not known, we use the standard error given by:

$$SE = \frac{S}{\sqrt{n}}$$

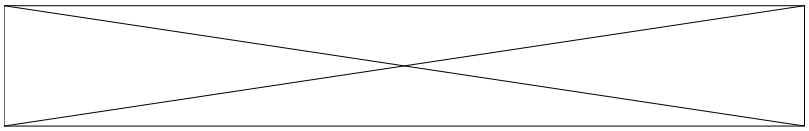
Where

$$S = \sqrt{\frac{(\sum x-\bar{x})}{n}} \quad \text{If n is large}$$

$$S = \sqrt{\frac{(\sum x-\bar{x})^2}{n-1}} \quad \text{If n is small}$$

In drawing statistical inferences, the standard error is of great significance due to

1. That it provides an idea about the reliability of sample. The lesser the standard error, the lesser the variation of population value from the expected (sample) value. Hence is greater reliability of sample.
2. That it helps to determine the confidence limits within which the parameter value is expected to lie. For large sample, sampling distribution tends to be close to normal distribution. In normal distribution, a range of mean ± one standard error, of mean±3 two standard error, of mean ± 3 standard error will give 68.27 per cent, 95.45 per cent and 99.73 per cent values respectively. The chance of a value lying outside ± 3 S.E. is only 0.27 per cent. i.e.. approximately 3 in 1000.
3. That it aids in testing hypothesis and in interval estimation.
4. That it aids in testing hypothesis and in interval estimation.



Please use headphones

Estimation Theory

A technique which is used for generalizing the results of the sample to the population for estimating population parameters along with the degree of confidence is provided by an important branch of statistics is called Statistical Inference. In other words, it is the process of inferring information about a population from a sample. This statistical inference deals with two main problems namely (a) estimation and(b) testing hypothesis.

a Estimation:

The estimation of population parameters such as mean, variance, proportion, etc., from the corresponding sample statistics is an important function of statistical inference. The parameters estimation is very much need for making decision. For example, the manufacturer of electric tubes may be interested in knowing the average life of his product, the scientist may be eager in estimating the average life span of human being and so on. Due to the practical and relative merits of the sample method over the census method, the scientists will prefer the former.

A specific observed value of sample statistic is called estimate. A sample statistic which is used to estimate a population parameter is known as estimator. In other words, sample value is an estimate and the method of estimation (statistical measure) is termed as an estimator. The theory, of Estimation was innovated by Prof. R.A. Fisher. Estimation is studied under Point Estimation and Interval Estimation.

Good Estimation

A good estimator is one which is as close to the true value of population parameter as possible. A good estimator possesses the features which are:

(a) Unbiasedness : An estimate is said to be unbiased if its expected value is equal to its parameter. For example, if 3c is an estimate of ft, x will be an unbiased estimate only if

i $E(\bar{x}) = \mu$

ii $E(S^2) = \theta^2$

(See Illustration 1)

(b) Consistency : An estimator is said to be consistent if the estimate tends to approach the parameter as the example size increases. For any distribution, i.e.., symmetrical or skew symmetric, sample mean, sample variance and sample proportions are consistent estimators of the population mean, population variance and population proportion respectively.

(c) Efficiency: An estimate is said to be efficient if the variance i.e., is minimum. An estimator with less variability and the consistency more reliable than the other.

(d) Sufficiency : An estimator which uses all the relevant information in its estimation is said to be sufficient. If the estimator sufficiently insures all the information in the sample, then considering the other estimator is absolutely unnecessary.

Point Estimation Method

A Point Estimation is a single statistic which is used to estimate a population parameter. Now, we shall discuss the sample mean and sample variance are unbiased estimate for corresponding population parameters.

Sample Mean and Sample Variance: A sample men a is the best estimator of population mea is unbiased, consistent and efficient estimator, and the sampling distribution is closer to no distribution if so long as sample is sufficiently large. The Central Limit Theorem tells that sampling distribution mean is equal to the population (3: . = *p)*. The variance of samp distribution is equation the population

variance divided by n $\left(s^2 = \frac{\theta^2}{n}\right)$ The standard deviation of the sampling distribution of a stat is known as its standard error. It is defined as

Interval Estimation Method

In Point Estimation, a single value of statistic is used as estimate of the population parameter. Sometimes, this point estimate may not disclose the true parameter value. Having computed a statistic from a given random sample, can we make reasonable probability statements about the unknown parameter of the population from which the sample is drawn? The answer can provide by the technique of Interval Estimation. The Interval Estimation within which the unkown value of parameter is expected to lie is called confidence interval or Fiducial Interval (which are respectively called by Neyman and Fisher). The limits so determined are called Confidence Limits or Fiducial Limits and at required precision of estimate say 95 percent is known as Confidence Coefficient. Thus, a confidence interval indicates the probability that the population parameter lies within a specified range of values. To compute confidence interval we require:

a. the sample statistic,
b. the standard error (SE) of sampling distribution of the statistic,

c. the degree of accuracy required (i.e. confidence coefficient) as reflected by the Z-value.

(a) In rswr,

$$\bar{x} \pm Z = (SE\bar{x})$$

$$SE\bar{x} = \frac{s}{\sqrt{n}}$$

(b) In rswor,

$$\bar{x} \pm Z = (SE\bar{x})$$

$$SE\bar{x} = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

(c) When S.D of population is not known

$$\bar{x} \pm Z = (SE\bar{x})$$

$$SE\bar{x} = \frac{s}{\sqrt{n-1}}$$

(d) For different means

$$\bar{x}1 - \bar{x}2 \pm Z\,SE(\bar{x}1 - \bar{x}2)$$

$$SE = \frac{S1^2}{n1} + \frac{S2^2}{n^2}$$

2. The Confidence Limits for Variance :

(a) Standard deviation (large samples)

$$S \pm X SE \text{ of } SD$$

$$SE = \frac{S}{\sqrt{2n}}$$

(b) Variance

Variance : $\pm Z(\text{SE of var})$

$$\text{SE of var} = S^2\sqrt{\frac{2}{n}}$$

(d)Coefficient of variance (C.V.)

$$C.V. \pm Z\ (\text{SE of C.V})$$

$$\text{SE of C.V} = \frac{\text{Var}}{2n}1 + \frac{2\text{var}^2}{10^4}$$

Z-Distribution

Interval estimation for large samples is based on the assumption that if the size of sample is large, the sample value tends to be very close to the population value. In other words, the size of sample is sufficiently large, the sampling distribution is approximately of normal curve shape. This is the feature of the central limit theorem. Therefore, the sample value can be used in estimation of standard error in the place of population value. The Z-distribution is used in case of large samples to estimate confidence limits. For small sample, instead of Z-values, t-values are studied to estimate the confidence limits. One has to know the degree of confidence level before calculating confidence limits. Confidence level means the level of accuracy required. For example, the 99 per cent confidence level means, the actual population mean lies within the range of the estimated values to a tune of 99 per cent. The risk is to a tune of one percent.

To find the Z-value corresponding to 99 per cent confidence level, divide that J confidence level by 2 i.e. 99/2 which gives 49.5, in probability terms it is 0.495. Identify this value in the Z-value. The Z-value corresponding to it can be identified in the left-most column and also in the top-most row. The confidence coefficient for 99 per cent confidence level is 2.51. The 99 per cent of items or cases falls within x ± 2.51 SE which means the sampling distribution will have 99 percent of the population.

Superiority of Interval Estimate

In estimating the value by the Point Estimate Method and the Interval Estimate Method, the former provides only a point in the sample with no tolerance or confidence level attached to it. The latter provides accuracy of the estimate at a confidence level. Further it helps in hypothesis testing and becomes a basis for decision-making under the conditions of uncertainly or probability. The interval estimate, therefore, has a superiority or practical application over the point estimate.

Illustration 1

A Universe consists of four numbers 3, 5, 7 and 9. Consider all possible samples of size two which can be drawn with replacement from the universe. Calculate the mean

and variance. Further, examine whether the statistics are unbiased for corresponding parameters. What is the sampling mean and sample variance?

Solution:

$$\mu = \frac{\sum x}{N}$$
$$= \frac{3+5+7+9}{4} = \frac{24}{4}$$
$$= 6$$

$$\text{Variance } \theta^2 = \left[(x-\mu)^2 + ...\right]N$$
$$= \frac{(3-6)^2 + (5-6)^2 + (7-6)^2 + (9-6)^2}{4}$$
$$= (3)^2 + (1)^2 + (1)^2 + (3)^2$$
$$= \frac{9+1+1+9}{4}$$
$$= \frac{20}{4}$$
$$= 5$$

Calculation of sample mean and sample variance

Any one of the four numbers, 3, 5, 7 and 9 drawn in the first draw can be associated with any one of these four numbers drawn at random with replacement in the subsequent draw i.e.. second draw. Hence, the total number of possible samples of size 2 is 4 ´ 4 = 16 , and is given by the cross product: (3, 5, 7, 9) ´ (3, 5, 7, 9) as shown below.

Sample No.	Sample Value	Sample mean	x - $E(\bar{x})$ ($\bar{x}$ = 6.0)	Var $\bar{x}$ = - $\Sigma(\bar{x})^2$
1	3,3	3.0	- 3.0	9.0
2	3, 5	4.0	- 2.0	4.0
3	3,7	5.0	- 1.0	1.0
4	3,9	6.0	0	0
5	5, 3	4.0	- 2.0	4.0
6	5,5	5.0	- 1.0	1.0
7	5,7	6.0	0	0
8	5, 9	7.0	+ 1.0	1.0
9	7,3	5.0	- 1.0	1.0
10	7,5	6.0	0	0
11	7,7	7.0	+ 1.0	1.0
12	7, 9	8.0	+ 2.0	4.0
13	9,3	6.0	0	0
14	9, 5	7.0	+ 2.0	1.0
.15	9,7	8.0	+ 2.0	4.0
16	9,9	9.0	+ 3.0	9.0
		96.0		40.0

Sample mean $\bar{x}$ $= \frac{\sum x}{n} = \frac{96}{16}$ 6.0

Sample Variance $= \frac{\sum x - E(\bar{x})^2}{n}$

= 250

Thus,

Sample mean = Population = 6.0

Sample variance = Population Variance =2.50

We conclude that the sample statistics of mean and of variance, and the corresponding population parameter are the same; Hence, the sample statistics are unbiased estimate for the corresponding population parameters.

Illustration 2

Consider a hypothetical three numbers 2, 5 and 8. Draw all possible samples of size 2 and examine the statistics are unbiased for corresponding parameters.

Solution

The given universe consists of three values namely 2,5,8. The total possible samples of size 2 is 3x3 = 9 and is given by cross product: (2,5,8) x (2,5,8). Thus there are 9 samples of size 2. They are shown in the} following table.

Sample No.	Sample Values	Sample mean	x-5	Var $\bar{x}=x-E(\bar{x})^2$
1	2,2	2.0	- 3.0	9.00
2	2,5	3.5	- 1.5	2.25
3	2,8	5.0	0	0
4	5,2	3.5	- 1.5	2.25
5	5,5	5.0	0	0
6	5,8	6.5	1.5	2.25
7	8,2	5.0	0	0
8	8,5	6.5	1.5	2.25
9	8, 8	8.0	3.0	9.00

Sample mean, $\bar{x}$ $= \frac{\sum x}{n}$

$= \frac{45}{9} = 5$

Sample variance, S2 $= \frac{x - \sum (\bar{x})^2}{n}$

$= \frac{27.00}{9}$

$= 3.00$

Population mean μ $= \frac{\sum x}{N}$

$= \frac{2+5+8}{3}$

$= \frac{15}{3} = 5$

Population variance θ^2 $= \sum \frac{(x - \Sigma)^2 + \ldots.}{N}$

$= \frac{(2-5)^2 + (5-5)^2 + (8-5)^2}{3}$

$= \frac{(3)^2 + (0)^2 + (3)^2}{3}$

$= \frac{19}{3} = 6.0$

Proof of unbiased:

(i) $\bar{x} = \mu = 5$

(ii) $S^2 = \frac{\theta^2}{Z} = \frac{6.0}{2} = 3.0$

Hence statistics are unbiased for corresponding parameters.

Illustration 3

Consider the population of 5 units with values 1,2,3,4 and 5. Write down all possible samples of size 2 without replacement and verify that sample mean is an unbiased

estimate of the population mean. Also calculate sampling variance and verify that (i) it agrees with the formula for variance of the sample mean and (ii) this variance is less than the variance obtained from the sampling with replacement (iii) and find the standard error.

Solution

$$\text{Population mean,} \quad = \mu = \frac{1+2+3+4+5}{5}$$

$$= \frac{15}{5} = 3$$

$$\text{Population variance } \theta^2 \quad = \sum \frac{(x-\Sigma)^2 + \ldots.}{N}$$

$$= (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2$$

$$= \frac{4+1+0+1+4}{5}$$

$$= \frac{10}{5} = 2$$

All possible samples of size 2 without replacement are given by Nc_n :

$$Nc_n = \frac{N!}{n!(N-n)!}$$

$$= \frac{5!}{2!(5-2)!}$$

$$= \frac{5!}{2! \quad 3!}$$

$$= \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} = 10$$

Sample mean variance

Sample No.	Sample values	Sample mean	x- 3.0	Variance $\bar{x} = (x - E(\bar{x})^2$
1	1, 2	1.5	-1.5	2.25
2	1, 3	2.0	-1.0	1.00
3	1, 4	2.5	-0.5	0.25
4	1, 5	3.0	0	0
5	2, 3	2.5	-0.5	0.25
6	2, 4	3.0	0	0
7	2, 5	3.5	0.5	0.25
8	3, 4	3.5	0.5	0:25
	3, 5	4.0	1.0	1.00
10	4,5	4.5	1.5	2.25
		30.0		7.50

Sample mean, $\bar{x} = \sum x/n$

$= 30/10 = 3.0$

Sample variance $\{ (S^2) = \sum (x - E(\bar{x}^2) + ...$

$= 7.50/10 \quad = 0.75$

Proof

i. $\bar{x} = \mu$, 3.0 = 3.0

Hence sample mean is an unbiased estimate of population mean.

Variance of sample (without replacement)

$$= \frac{N-n}{N-1} \times \frac{\theta^2}{N} \times 1/2$$

$$= \frac{5-2}{5-1} \times \frac{10.0}{5} \times 0.5$$

Se (rswr) = $\sqrt{100}$ = 1.000

Thus, the variance of sample mean distribution is agreed with the formula for the variance of the sample without replacement.

Illustration 4

The Golden Cigarettes Company has developed a new blended tobacco product. The marketing, department has yet to determine the factory price, A sample of 100 wholesalers were selected and were asked about price. Determine the sample mean for the following prices supplied by the wholesalers.

Rs. 4.90 5.10 5.40 5.80

5.00

4.90 6.30 6.10 5.70

5.20

Solution

Calculation of sample mean, variance and standard deviation

Sample price (x)	Sample mean ($\bar{x}$)	$(x - \bar{x})$	$(x-\bar{x})^2$	X^2
4.90	5.50	- 0.60	0.36	24.01
5.70	5.50	+ 0.20	0.04	32.49
5.40	5.50	- 0.10	0.01	29.16
5.80	5.50	+ 0.30	0.09	33.64
5.00	5.50	- 0.50	0.25	25.00
4.90	5.50	-0.60	0.36	24.01
6.30	5.50	+ 0.80	0.64	39.69
6.10	5.50	+ 0.60	0.36	37.21
5.70	5.50	+ 0.20	0.04	32.49
5.20	5.50	- 0.30	0.09	27.04
	55.00		2.24	304.74
Sample mean	= 55/10 = 5.5			
Sample variance	= 2.24/10-1 = 2.24/9			

Standard deviation

$$S = \frac{\sum x^2 - (\bar{x})^2 \times n}{n-1}$$

$$= \frac{304.74 - (5.5)^2 \times 10}{10-1}$$

$$= \frac{2.24}{9}$$

$$= 0.2489 \quad \text{(by direct method)}$$

$$S = \frac{(-\bar{x})^2}{n-1}$$

$$= \frac{2.24}{10-1}$$

$$= 0.2489 \quad \text{(by Short-cut method)}$$

Thus, using the sample wholesaler's cigarettes price mean as an estimator, the point estimation of the golden cigarettes mean is Rs. 5.50. Both the buyer as well as seller accept the use of this point estimate as a basis for fixing the price. The point estimate can save time and expense to the producer of cigarettes.

Illustration 5

Sensing the downward in demand for a product, the financial manager was considering shifting his company's resources to a new product area. He selected a sample of 10 firms in the textile industry and discovered their earnings (in %) on investment. Find point estimate of the mean and variance of the population from data given below.

18.0	25.0	13.0	21.0	17.0
16.0	12.0	10.0	20.0	

Solution

Calculation of point estimate of sample mean and its variance

X	$\bar{x}$	$(x-\bar{x})$	$(x-\bar{x})^2$
18.0	16.0	+ 2.0	4
25.0	16.0	+ 9.0	81
13.0	16.0	- 3.0	9
11.0	. 16.0	- 5.0	25
21.0	16.0	+ 1.0	25
17.0	16.0	0	1
16.0	16.0	- 4.0	16
12.0	16.0	-6.0	36
17.0	16.0	+ 1.0	1
160.0			198
Sample mean		$=160/10=16$	
Sample variance		$=198/10-1=198/9=22$	
Sample standard deviation		$=\sqrt{22}=4.69$	

Thus, the point estimate of mean and of variance of the population from which the sample drawn are 16 and 4.69 respectively.

Thus, the point estimate of mean and of variance of the population from which the sample drawn are 16 and 4.69 respectively.

Illustration 6

A random sample of 600 appeals was taken from a large consignment and 66 of them were found to be bad. Find the limits at which the bad appeals lie at 99 per cent confidence level.

Solution

Calculation of confidence limits for the proportion of bad appeals.

We are given,

$$N=600$$

Number of bad appeals in the consignment, 65 proportion of bad appeals in the consignment,

$$P = a/n = 66/600 = 0.11$$

proportion of good appeals in the consignment,

$$q = 1-\frac{a}{n} = 1.00-0.11 = 0.89$$

Standard error of proportion of bad appeals in consignment is given by:

$$SE(p) = \sqrt{pq/n}$$

$$= \sqrt{\frac{0.11\times 0.89}{600}}$$

$$= 0.0128$$

The value of statistic at 99,73 per cent confidence level is 3 which is tested to find the most probable limits within which bad appeals lie :

$$p \pm Z\sqrt{pq/n}$$

$$p \pm 3\sqrt{pq/n}$$

$$0.11 \pm 3\times 0.0128$$

$$0.11 \pm 0.0384$$

0.1484 and 0.0716

Hence, the bad appeals in the consignment lie between the limits at 14.84 per cent and 7.16 percent.

Illustration 7

Out of 20,000 customer's ledger accounts, a sample of 500 accounts was taken to accu-racy of posting and balancing wherein 40 mistakes were found. Assign limits Within which the number of defective cases can be expected to lie at 95 per cent confidence.

Calculation of confidence limits for defective cases.

We are given,

N= 500 and N = 20,000

No. of mistakes, x= 40

Therefore,

P = x/n =40/500 =0.08

q = 1.00-0.08 =0.92

Estimation of error $SE(p) = \sqrt{pq/n}$

$$= \sqrt{\frac{0.08 \times 0.92}{500}}$$

$= 0.0121$

95 per cent confidence limits for proportion of mistakes is given by (95% confidence value = 1.96)

$p \pm 1.96 \times 0.0121$

0.08 ± 0.237

0.1037 and 0.0563

Hence, the number of mistakes in a lot of 20,000 are expected to lie are (2000 x 0.1037) 2024 and (20000 X 0.0563) 1126.

Remark

The universe, N is sufficiently large to the sample size, n, the sampling fraction n/N is to be taken into account. Accordingly as a sample from large population using rswor, an estimate of sample is 0.08 (as above). By using the 95 per cent confidence limits for proportion of mistakes are:

$$p \pm 1.96 \quad \frac{\sqrt{(N-n)pq}}{N(n-1)}$$

$$0.08 \pm 1.96 \quad \frac{\sqrt{(N-n)pq}}{N(n-1)}$$

$$0.08 \pm 1.96 \quad \frac{\sqrt{20000-500 \times 0.08 \times 0.92}}{20000(500-1)}$$

$$0.08 \times 1.96 \quad \frac{\sqrt{19500 \times 0.08 \times 0.92}}{20000 \times 499}$$

$0.08 \pm 1.96 \times 0.01199$

0.08 ± 0.0235

0.1035 and 0.0565

Hence, the required number of defective cases in the lot lies between:

20000 (0.1035, 0.565)

i.e. 2070 and 1130

NOTE

It is proved that when the sample size is large the sample value is tends to be close to the population value.

Illustration 8

In sample of 1000 TV viewers, 330 watched a particular programme. Find 99 per cent confidence limits for TV viewers who watch this programme.

Solution

In notation

n = 1000 and x = 330

p(proportion of TV viewers) = x/n = 330/1000 = 0.33

Then, q=1-p = 1-0.33 = 0.67

$$SE(p) = \sqrt{pq/n} = \frac{\sqrt{0.33 \times 0.67}}{1000}$$

$$= 0.0149$$

Z at 99 per cent confidence value = 2.58

$p \pm 2.58\, SE\,(p)$

$0.33 \pm 2.58 \times 0.0149$

0.33 ± 0.03844

0.3684 and 0.2916

TV viewers are 36.84 per cent and 29.16 per cent. The number of TV viewers in 1000 lies between 1000 (0,2916 and 0.368 i.e. 292 and 365.

Illustration 9

Out of 1200 M.Com. students, a, sample of 150 selected at random to test the accuracy of solving a problem in Quantitative Methods and 10 did mistakes. Assign limits within which the number of students who done the problem wrongly in whole universe of 1200 students at 99 per cent confidence level.

Solution

We are given

N = 1200, n = 150 and x = 12

Proportion of students who did the problem,

P= 12/150 = 0.08

Proportion of students does the problem,

q = 1-p = 1.00 -0.08 = 0.92

$$SE(p) = \sqrt{pq/n} = \frac{\sqrt{0.08 \times 0.92}}{150}$$

$= 0.022$

$p \pm z\ SE\ (p)$

$0.08 \pm 2.58 \times 0.022$

0.08 ± 0.05676

0.1368 and 0.2324

The number of students who did the problem wrong is laid between 150 (0.02324, 0.1368) i.e. 3 and 20.

The number of students who did wrong the problem in the whole universe 1200 at 99 per cent level is:

$$P \pm 2.58 = \frac{\sqrt{1200-150 \times 0.08 \times 0.92}}{1200\,(150-1)}$$

$$0.08 \pm 2.58 = \frac{\sqrt{1050 \times 0.08 \times 0.92}}{1200 \times 149}$$

$$0.08 \pm 2.58 = \frac{\sqrt{72.28}}{178800}$$

$$0.08 \times 2.58 \times 0.02079$$

$$0.08 \pm 0.05364$$

0.13364 and 0.02636

Thus, the probable, percentage of students did the problem wrongly in the whole universe of 1200 lied in between 13.36 and 2.64. And the number of students were in between 160 and 32.

- End Of Chapter -

LESSON - 10

HYPOTHESIS

To be fruitful, the decision, one should collect random Sample for or against some point of view of proposition. Such point of view or proposition is termed as hypothesis. Hypothesis is a proportion which can be put to test to determine validity. A hypothesis, in statistical parlance is a statement about the nature of a population which is to be tested on the basis of outcome of a random sample.

Testing Hypothesis

The testing hypothesis involves five steps which are as:

(a) Hypothesis Formulation:

The formulation of a hypothesis about population parameter is the first step in testing hypothesis. The process of accepting or rejecting a null hypothesis on the basis of sample results is called testing of hypothesis. The two hypothesis in a statistical test are normally referred to:

i. Null hypothesis

ii. Alternative hypothesis.

A reasoning for possible rejection of proposition is called null hypothesis. In other words, it asserts that there is no true difference in the sample and the population, and that the difference found is accidental and unimportant arising out of fluctuations of sampling. Hence the word, null means invalid, void or amounting to nothing. Decision-maker should always adopt the null attitude regarding the outcome of the sample.

A hypothesis is said to be alternative hypothesis when it is complementary to the null hypothesis. The null hypothesis and alternative hypothesis are denoted by Ho and Hi respectively.

(b) Type I and Type II errors:

A null hypothesis consists of only a single parameter value and is simple while the alternative hypothesis is usually composite. In any statistical test, there are four possibilities which are termed as exhaustive decisions.

They are:

1. Reject H_0 when it is false
2. Accept H_0 when it is true
3. Reject H_0 when it is true (Type -I error)
4. Accept H_0 when it is false (Type - II error)

The decisions are expressed in the following dichotomous table:

	Accept H_o	Reject H_o
H_o True	Correct decision	Wrong decision Type I error
H_o False	Wrong decision Type II error	Correct decision

The error of rejecting Ho when it is true is called Type I error and of accepting Ho when Ho is false is known as Type II error. The probability denoted by a (pronounced as alpha) and the probability (p) of type II error is denoted by a (pronounced as beta). In practice, in business and social science problems, it is more risky to reject a correct hypothesis than to accept a wrong hypothesis. In other words, the consequences of Type I error are likely to be more serious than the consequences of Type II error.

(c) Level of Significance:

The quantity of risk tolerated in hypothesis testing is called the level of significance is commonly used at 5 percent respectively account for moderate and high precision.

(d) Test Statistics:

The most commonly used test are t-test, F-ratio and Chi-square. The estimated value of the parameter which depends on the number of observations. The sample size, therefore, plays an important role in testing of hypothesis and is taken care of by degrees of freedom. Degrees of freedom are the number of independent observations in a set.

(e) Conclusion:

A statistical decision is a decision either to accept or to reject the null hypothesis based on the computed value in comparison with the given level of significance. If the computed value of test statistic is less or more than the critical value, it can be said

that the significant difference or insignificant difference and, the null hypothesis is rejected or accepted respectively at the given level of significance.

Test of Significance

The tests of significance available to know the significance or otherwise of variables in various situations are

(a) Test of significance for large samples and

(b) Test of significance for small samples.

These tests aim at

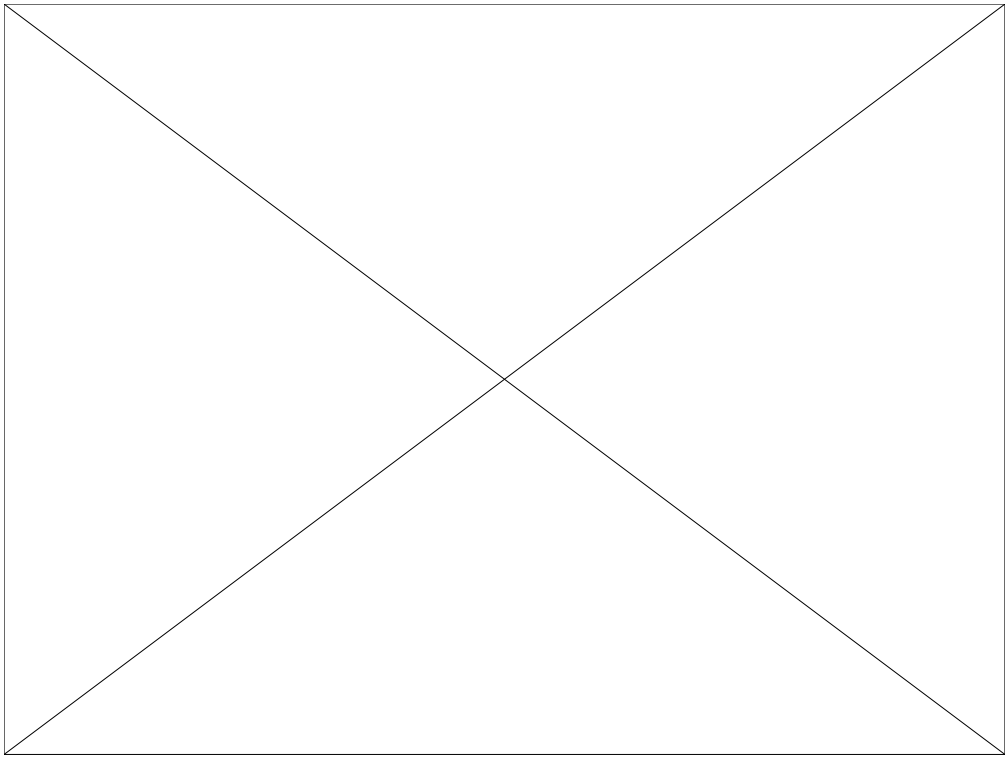
(i) comparing observation with expectation and thereon finding how far the deviation of one from the other can be attributed to variations of sampling,

(ii) estimating from samples some characteristic of the population and

(iii) guaging the reliability of estimates.

It is difficult to draw a line of demarcation between large and small samples; but a view among statistician is that, a sample is to be recorded as large only if its size exceeds 30 and if the sample size is less than 30, it is noted as small sample. The tests of significance used for large samples are different from the small samples, the reason being that the assumptions made in case of large samples do not hold good for small samples. The assumptions that made in dealing with problems relating to large samples are: (a) the random sampling distribution of a statistic is approximately normal and (b) the values given by sample data are sufficiently close to the population values and can be used in their (population), place for calculating the standard error of the estimate.

In case of small samples, the above said assumptions will no longer be hold good. It should be noted that the estimates will vary from sample to sample if we work with very small samples. We must satisfy with relatively wide confidence intervals. Of course, the wider the interval, the less is the precision. An inference drawn from the large sample is far more precise in the confidence limits it sets up than an inference based on a much smaller sample. Though, drawing a precise line of demarcation between the large sample and the small sample is not always easy, but the division of their theories is a very real one. As a rule, the theory and methods of small samples are applicable to large samples, but the reverse is riot true.



Please use headphones

Large Samples

(a) Single mean:

Illustration 10

Compute the standard error of mean from the following data showing the amount mid by 100 firms on the occasion of Deepavali.

Amount paid (Rs.)	No. of firms
30-40	2
40-50	4
50-60	10
60-70	20
70-80	34
80-90	25
90-100	5

Solution

Formula, SE $\bar{x}$ = **s /** $\sqrt{n}$

Amount Rs.	Mid value X	X = x-65/10	f	fdx	f.dx^2
30-40	35	- 3	2	- 6	18
40-50	45	-2	4	- 8	16
50 60	55	- 1	10	- 10	10
00-70	65	0	20	0	0
70-80	75	+ 1	34	+ 34	34
80-90	85	+ 2	25	+ 50	100
90-100	95	+ 3	5	+ 15	45
			100	75	223

$$S = \sqrt{\frac{\sum fdx^2}{n} - \left(\frac{\sum fdx}{n}\right)^2} \times c$$

$$= \sqrt{\frac{223}{100} - \left(\frac{75}{100}\right)^2} \times 10$$

$$= \sqrt{2.23 - 0.5625} \times 10$$

$$= 12.91$$

$$SE\bar{x} = 12.91 / \sqrt{100}$$

$$= 1.291$$

Illustration 11

For a random sample of 100, the mean height is 63 inches. The standard deviation of the height distribution of the population is known to be 3 inches. Test the statement that the mean height of the population is 66 inches at 0.05 level of significance. Also set up 0.01 confidence limits of the mean height of the population.

Solution

Let H_o: $\bar{x} = \mu$

SE $\bar{x} = S/\sqrt{n}$

$= 3/\sqrt{100}$

Difference $= \bar{x} - \mu = 63-66 = 3$

$$\text{Expected mean height of population} = \frac{\text{Difference}}{\text{SE}}$$

$= 3/0.3$

$= 10.00$

Since the difference is more than 1.96 SE at 0.05 level, it could not arise due to fluctuations of sampling. Hence, the hypothesis is rejected. In other words, the mean height of the population could not be 66 inches.

99 percent confidence limits of the mean height of the population,

$\bar{X} \pm Z$ SE

$63 \pm 2.58\ (0.3)$

63 ± 0.774

62.20 and 63.80

Illustration 12

A sample of 900 items is taken from a normal population whose mean is 6 and variance is 6. If the sample mean is 6.60 can the sample be regarded as a truly random sample. Give necessary justification for your conclusions.

Solution

Let us take the hypothesis that there is no difference between the sample mean a the population mean.

$$SE\ \bar{x} = S/\sqrt{n}$$

$$S = \sqrt{Variance}$$

$$= \sqrt{6}$$

$$= 2.45$$

$$SE\ \bar{x} = 2.45/\sqrt{900}$$

$$= 0.0817$$

Difference between the sample mean and population mean

6.60 - 6.00 = 0.60

$$Z\ (\text{Expected mean of population}) = \frac{\text{Difference}}{SE}$$

$$= \frac{0.60}{0.0817} = 7.34$$

Since the difference is more than 1.96 SE at 0.05 level, it could not have arisen due to variations of sampling. Hence, the sample cannot be regarded as truly random sample.

Illustration 13

If it costs a rupee to draw one number of a sample, how much would it cost in sampling from a universe with mean 100 and standard deviation 9 to take sufficient number as to ensure that the mean of a sample would be within 0.015 per cent of the value at 0.05 level. Find the extra cost necessary to double this precision.

Solution

We are given,

x = 100 and S.D = 9

The difference between sample mean and population mean = 0.015 (given)

For 95 percent confidence, difference between sample mean and population mean should be equal to 1.96 SE.

0.015 = 1.96 SE

$$SE = s/\sqrt{n}$$

Therefore,

$$1.96\ S/\sqrt{n} = 0.015$$

$$1.96 \times \frac{9}{\sqrt{n}} = 0.015 = 1.96 \times 9 = 0.015 \times \sqrt{n}$$

$$\sqrt{n} = \frac{1.96 \times 9}{0.015}$$

$$\sqrt{n} = (1176) \quad \text{(by squaring both sides)}$$

$$n = (1176)^2$$

$$= 13, 82,976$$

The difference between sample mean and population mean by making the precis

le, it should be = 0.015 / 2.0 = 0.0075. So,

$$\frac{1.96 \times 9}{\sqrt{n}} = 0.0075$$

$$\sqrt{n} = \frac{1.96 \times 9}{0.0075} \quad \text{(by squaring both sides)}$$

$$n = 55,31,904$$

Therefore,

Extra cost = 55,31,904 – 13,82,976

= Rs. 41,48,928

Hence to double the precision, the extra cost would be Rs. 41, 48,928.

Illustration 14

The average number of defective articles in a factory is claimed to be less than for all the factories whose average is 30.5. A random sample showed the following distribution.

Class Limits	Number

16-20	12
21-25	22
26-30	20
31-35	30
36-40	16

Calculate the mean and standard deviation of the sample and use it to test the claim that the average is less than the figure for all the factories at 0.05 level of significance.

Solution

The sample mean and population mean do not

i.e., H_o: $\bar{x} = \mu$

Class limits	Mid value	dx= x-28/5	f	fdx	$f.dx^2$
16-20	18	-2	12	-24	48
21-25	23	-1	22	-22	22
26-30	28	0	20	0	0
31-35	33	+1	30	+30	30
36-40	38	+2	16	+32	64
			100	16	164

$$\bar{x} = A + \frac{fdx \times c}{n}$$

$$= 28 + \frac{16 \times 5}{100}$$

$$= 28 + 0.8 = 28.8$$

$$S = \sqrt{\frac{\sum fdx^2}{n} - \left(\frac{\sum fdx}{n}\right)^2} \times c$$

$$= \sqrt{\frac{164}{100} - \left(\frac{16}{100}\right)^2} \times 5$$

$$= \sqrt{1.64 - 0.0256} \times 5$$

$$= \sqrt{1.6144} \times 5$$

$$= 6.35$$

Test statistic

$$Z = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$= \frac{28.8 - 30.5}{6.35/\sqrt{100}}$$

$$= \frac{-1.7\ \sqrt{100}}{6.35}$$

$$= \frac{-17.00}{6.35}$$

$$= -2.68$$

Since f |Z| is more than 1.96, it is signification at 0.05 level of significance. Hence we reject the null hypothesis and conclude that the sample mean and population mean differ significantly. In other words, the manufacturer's claim that the average

number of defectives in his product is less than the average figure for all the factories is valid.

Illustration 15

A random sample of 100 articles selected from a batch of 2000 articles which show that the average diameter of the articles is 0.354 with a standard deviation 0.048. Find 95 per cent confidence interval for the average of this batch of 2000 articles.

Solution

We are given,

$$n = 100, \quad N = 2000, \quad \bar{x} = 0.354 \quad \text{and} \quad S = 0.048$$

$$SE\ \bar{x} = \frac{s}{\sqrt{n}} \times \frac{N-n}{N-1} \text{ (Sample wihtout replacement)}$$

$$= \frac{0.048}{\sqrt{100}} \times \frac{2000-100}{2000-1}$$

$$= 0.0048 \times 0.9749$$

$$= 0.004668$$

95 percent confidence limits for population mean is

$$\bar{x} \pm Z\ SE\bar{x}$$

$$\bar{x} \pm 1.96\ (0.00468)$$

0.354 X 0.0092

0.3448 and 0.3632

Illustration 16

A sample of 400 male students is found to have a mean height of 171.38 cms. Can it reasonably be regarded as a sample from a large population with mean height of 171.17 cms. and standard deviation 3.30 cms?

Solution

HO = m = 171.17

We are given,

$H_0 = \mu = 171.17$

We are given,

$\bar{x} = 171.38, \quad n = 400, \quad S = 3.30 \quad \text{and} \quad \mu = 171.17$

Test statistics

$$Z = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$= \frac{171.38 - 171.17}{3.30/\sqrt{400}}$$

$$= \frac{0.21 \times \sqrt{400}}{3.30}$$

$$= 1.273$$

Since |Z| is less than 1.96 at 0.05 level of significance, we accept the null hypothesis. In other words, the sample of 400 has come from the population with mean height of 171.17.

Illustration 17

Mrs. P, an insurance agent in Anantapur Division has claimed that the average age of policy-holders who insure through her is less than the average of all the agents, which is 30.5 year. A random sample of 60 policy-holders who had insured through her gave the following age distribution.

Age last birthday	No. of person
16-20	4
21-25	3
26-30	15
31-35	18
36-40	10

Calculate the mean and standard deviation of this distribution and use these values to test her claim at the 95 per cent of level of significance. You are given that Z at 0.95 is 1.96.

Solution

Age	Mid value	dx = (x - 28)/5	f	fdx	f.dx²
16-20	18	- 2	4	- 8	16
21-25	23	- 2	13	-13	13
26-30	28	0	15	0	0
31-35	33	+ 1	18	+ 18	18
36-40	38	+ 2	10	+ 20	40
			60	17	87

Mean $= A + \frac{fdx}{N} \times c$

$= 28 + \frac{17}{60} \times 5$

$= 28 + 1.42$

$= 29.42$

S $= \sqrt{\frac{\sum fdx^2}{N} - \left(\frac{\sum fdx}{N}\right)^2} \times c$

$= \sqrt{\frac{87}{60} - \left(\frac{17}{60}\right)^2} \times 5$

$= 5.85$

Z $= \frac{\bar{x} - \mu \times \sqrt{n}}{s}$

$$= \frac{29.42 - 30.50 \times \sqrt{60}}{5.85}$$

$$= 1.43$$

Since | Z | is less than 1.96 at 0.05 level of significance, it could have arisen due to sampling variation. Hence, difference is insignificant. In other words, Mrs. P claim that the average age of policy-holders who insure through her is less than the average for all the agents at 30.5 years, is valid,

(b) Two Means:

Illustration 18

A random sample of 1000 workers from South India shows that their mean wage of Rs. 47 per week with a standard deviation of Rs. 28. A random sample of 1500 workers from North India gives a mean wage of Rs. 49 per week with a standard deviation of Rs. 40. Is there any significant difference between their mean wages?

Solution

We are given two independent samples. Their means be denoted by *x* and ; sizes by m and n2 respectively. The given values are:

$n_1 =$ 1000 $\bar{x}_1 = 47$, $S_1 = 28$

$n_2 =$ 1500 $\bar{x}_2 = 49$, $S_2 = 40$

Let Ho: $\mu_1 = \mu_2$

Formula

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Substituting the value, we get

$$Z = \frac{47 - 49}{\sqrt{\frac{28^2}{1000} + \frac{40^2}{1500}}}$$

$$= \frac{-2}{\sqrt{0.784 + 1.067}}$$

$$= -2/1.361$$

$$= 1.47$$

Since |Z| is 1.47 which is less than 1.96, the z value is not significant at 0.05 level of signi
This implies that the data do not provide any evidence against null hypothesis, which may, theref
accepted. We conclude that μ_1 is equal to μ_2 i.e. the mean wages do not differ significantly in Sou
India and North India.

Illustration 19

Mrs. Lahari has selected two markets, A and B at different locations of a city in order to make a survey on buying habits of customers. 400 women shoppers are chosen at random in market A. Their average monthly expenditure on food is found to be Rs. 1050 with a standard deviation of Rs. 44. The corresponding figures are Rs. 1020 and Rs. 56 found respectively in market B where also 400 men shoppers are chosen at random. Test at 1 per cent level of significance whether the average monthly food expenditure of the two populations of shoppers are equal.

Solution

Let H_o: $\mu 1 = \mu 2$

Given values are,

$n_1 =$ 400, $\bar{x}_1 = 1050$, $S_1 = 44$

$n_2 =$ 400, $\bar{x}_2 = 1020$, $S_2 = 56$

Test statistic

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$Z = \frac{1050 - 1020}{\sqrt{\frac{44^2}{400} + \frac{56^2}{400}}}$$

$$= \frac{30}{3.56} = 8.43$$

Since | Z | is 8.43 which is much greater than 2.58, the value of Z at 1 per cent level of significance, it is highly significant. Hence, the data do not provide any evidence to accept the hypothesis. In other words, the monthly expenditure in two populations of shoppers in markets A and B differ significantly

Illustration 20

An examination was given to 50 students at College A and to 60 students at College B. At A, the mean grade was 75 with a standard deviation of 9, at B, the figures were 79 and 7 respectively. Is there a significant difference between the performance of the students at A and at B, given at 0.05 and 0.01 level of significance?

Solution

Let the hypothesis that there is no significant difference between the performance of the students at college A and B.

We are given the value of

$n_1 = 50, \ \bar{x}_1 = 75, \ S_1 = 9$

$n_2 = 60, \ \bar{x}_2 = 79, \ S_2 = 7$

Test statistics

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$= \frac{75 - 79}{\sqrt{\frac{(9)^2}{50} + \frac{(7)^2}{60}}}$$

$$= -2.56$$

Conclusion

i. At 0.05 level of significance, since |Z| is 2.56 which is greater than 1 96, it is significant difference and therefore, we reject the hypothesis. In other words, we conclude that mean grades of the students of college A and B are different at 0.05 level of significance.

ii. At 0.01 level of significance - The value | Z | 2.56 which is less than the value of Z at 1 percent level of significance i.e. 2.58. Thus, the data consistent with the hypothesis and conclude that the mean grades of the students of college A and B is almost the same.

Illustration 21

Random samples drawn from two States give the following data relating to the heights of adult males.

	State A	State B
X	67.42	67.25
S	2.58	2.50
n	1000	1200

Is the difference between the means significant?

Solution

Let the hypothesis be there is no significant difference in mean height of adult males of two States.

Given values are:

$n_1 = 1000, \quad \bar{x}_1 = 67.42, \quad S_1 = 2.58$

$n_2 = 1200, \quad \bar{x}_2 = 67.25, \quad S_2 = 2.50$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$= \frac{67.42 - 67.25}{\sqrt{\frac{(2.58)^2}{1000} + \frac{(2.50)^2}{1200}}}$$

$$= \frac{0.17}{0.109} = 1.56$$

Since |Z| value is less than 1.96 at 0.05 level of significance, we accept the hypothesis.

(c) Single Variance:

Illustration 22

A sample survey of 121 boys about their intelligence gives a mean of 84, with-a standard deviation of 10. The population standard deviation is 11. Does the sample has come from the population ?

Solution

Sample has come from the population

$$Z = \frac{S-0}{S/\sqrt{2n}}$$

$$= \frac{10-11}{10/\sqrt{2\times 121}}$$

$$= \frac{1}{10\times\sqrt{2\times 121}}$$

$$= 1.55$$

Since |Z | is less than 1.96 at 0.01 level of significance, we conclude that the sample with standard deviation of 10 has come from the population.

(d) Two Variance :

Illustration 23

The mean yield of two sets of plots and their variability are given below. Examine whether the difference in the variability of yields is significant.

	Set of 40 plots	set of 60 plot
X	500 kgs	492kgs
S2	26kgs	20kgs

Solution

Let the hypothesis be that there is no significant difference in the variability of yield.

Given values are:

$n_1 =$ 40, $\bar{x}_1 = 500$, $S_1 = 29$

$n_2 =$ 60, $\bar{x}_2 = 492$, $S_2 = 20$

Substituting the values in the test statistic,

$$Z = \frac{2.58 - 2.50}{\sqrt{\frac{(2.58)^2}{2 \times 1000} + \frac{(2.50)^2}{2 \times 1200}}}$$

$= 0.8/0.077$

$= 10.39$

Since |Z| value is much greater that the 2.58 at 0.01 level of significance, it could not have arisen due to variations of sampling. We may reject the hypothesis.

Illustration 24

A sample of height of 6400 Englishmen has a mean of 67.85 inches and a standard deviation of 2.56 inches. While a sample of height of 1600 Australians has a mean 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that Australians are, on an average, taller than Englishmen?

Solution

(i) $H_0 : S1 = 2s$, Given values are

$n_1 =$ 6400 $\bar{x}_1 = 67.75$, $S_1 = 2.56$

$n_2 =$ 1600 $\bar{x}_2 = 68.55$, $S_2 = 2.52$

Estimation of Z value

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{S_1^2}{2n_1} + \frac{S_2^2}{2n_2}}}$$

$$= \frac{2.56 - 2.52}{\sqrt{\frac{(2.56)^2}{2\times 6400} + \frac{(2.52)^2}{2\times 1600}}}$$

$$= 0.80$$

Since |Z| is 0.08 which is less than 1.96 at 0.05 level of signifcance. Hence accept the hypothesis, i.e. there is no significant difference in variability of height between Englishmen and Australian.

(ii) Ho : $\mu_1 = \mu_2$, Given values are

$n_1 =$ 6400, $\bar{x}_1 = 67.75$, $S_1 = 2.56$

$n_2 =$ 1600, $\bar{x}_2 = 68.55$, $S_2 = 2.52$

Estimation of Z value

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$= \frac{67.85 - 68.55}{\sqrt{\frac{(2.56)^3}{6400} + \frac{(2.52)^2}{1600}}}$$

$$= \frac{0.7}{0.071}$$

$$= 9.86$$

Since |Z| is 0.08 which is much greater than 1.96 at 0.05 level of significance. Hence reject the hypothesis.

- End Of Chapter -

LESSON - 11

SMALL SAMPLES

If sample size is less than 30, it is termed as small sample. The greatest contribution to the theory of small samples is that of Sir William Gossett (t-test), R.A. Fisher (F-Test), Karl Pearson (Chi-Square test).

(a) t-test

The t-test will be studied under the assumptions namely:

(i) the parent population from which the sample is drawn is normally distributed,

(ii) the sample observations are random and independent of each other and

(iii) the population standard deviation is unknown.

The t-test is to be applied to test:

(a) the significance of single mean,

(b) the significance of the two independent samples, and

(c) the significance of the two dependent samples.

Single Mean

We calculate the statistics for determining whether the mean of a sample drawn from a normal population deviates significantly from the stated value (hypothetical population mean) of the statistics is defined as :

$$t = \frac{\bar{x} - \mu}{S} \times \sqrt{n}$$

$$S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}} \quad \text{When} \quad x - \bar{x} = 0$$

If calculated 111 is more than the tabulated t for n-1 degrees of freedom at certain level of significance, we say it is significant and Ho is rejected. If calculated 111 is less than tabulated t, H_o may be accepted at the adopted level of significance.

Illustration 25

A machine is designed to produce insulating washers for electrical devices of average thickness of 0.025 cms. A random sample of 10 washers was found to have an average thickness of 0.024 cms with a standard deviation of 0.002 cms. Test the significance of the deviation.

Solution

We are given,

x = 0.024 cms, n = 10, S = 0.002 cms and

m = 0.025 cms.

$$= \frac{0.024 - 0.025}{0.002} \times \sqrt{10}$$

$$= \frac{0.001 \times 3.16}{0.002} = 1.58$$

Since |t| is 1.58 which is less than the tabulated value at 9 d.f at 0.05 level of significance. Hence, deviation is not significant. H_o is accepted.

Illustration 26

A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of squares of deviations from this mean equal to 135 inches. Show that the assumption of a mean of 43.5 inches for the population is not reasonable. Obtain 95 per cent confident limits for the same.

Solution

Ho : No significan t difference in means between sample and population .

We are given $n = 16$, $\Sigma(x - \bar{x})^2$, = 135, $\bar{x}$ = 41.5 and μ = 43.5

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{(E(x - \bar{x})^2}{n-1}}} \times \sqrt{n}$$

$$= \frac{41.5 - 43.5}{\sqrt{\frac{135}{16-1}}} \times \sqrt{16}$$

$$= \frac{2.0 \times 4}{3}$$

$$= 2.67$$

Since |t| is more than the table value of 2.131 at t 0.05 for 15 d.f Hence HO is rejected. We conclude that the assumption of a mean of 43.5 inches for population is not reasonable.

$$SE\ \bar{x} = s/\sqrt{n}$$

$$S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

$$S = \sqrt{135/16-1}$$

$$= 3$$

$$SE\ \bar{x} = 3/\sqrt{16}$$

$$= 0.75$$

95 per cent fiducial lim its for population mean $\bar{x} \pm 0.05$ at 15 d.f. $(SE\ \bar{x})$ is

41.5 ± 2.131 ´ 0.75

41.5 ± 1.598

39.902 and 43.098

Therefore,

39.902 and 43.098

Illustration 27

Ten individuals are chosen at random from a population and their heights (in inches) are found to be: 62, 63, 66, 68, 69, 71, 70, 68, 71 and 66. In the light of these data, mentioning the null hypothesis, discuss the suggestion that the mean height in the population is 66 inches.

Solution

H_o : μ= 66 inches

X	dX = x - 67	dX^2
62 63 66 68 69 71 70 68 71 66	-5 - 4 - 1 1 2 4 3 1 4 - 1	25 16 1 1 4 16 9 1 -16 1
	4	90

$$\bar{x} = A + \frac{\Sigma dx}{n}$$
$$= 67 + \frac{4}{10}$$
$$= 67.4$$

$$S = \sqrt{\frac{\Sigma dx^2 - \left(\frac{\Sigma dx}{n}\right)^2}{n-1}}$$
$$= \sqrt{\frac{90 - \frac{(4)^2}{10}}{10-1}}$$
$$= \sqrt{\frac{90 - \frac{16}{10}}{10}}$$
$$= \sqrt{\frac{88.4}{9}}$$
$$= 3.13$$

$$t = \frac{\bar{X} - \mu}{S}\sqrt{n}$$
$$= \frac{67.4 - 66}{3.13} \times 3.16$$
$$= 1.41$$

Table value for 9 d.f at $t_{0.05}$ is 2.262 which is more than the calculated. It is not significant. Hence hypothesis at 0.05 level of significance may be accepted.
We conclude that the mean height in the population may be regarded as 66 inches.

Two Independent Samples Means

Given two independent random samples of size n_1 and n_2 with $\bar{x}_1$ and $\bar{x}_2$, and standard deviations S_1 and S_2, we can calculate the statistic 't' to test whether the samples have come from the normal populations. The statistic t is defined.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

S = Combined standard deviation

$$S = \sqrt{\frac{\Sigma(x_1 - \bar{x}_1) + \Sigma(x_2 - \bar{x}_2)}{n_1 + n_2 - 2}} \quad when \quad x - \bar{x} = 0$$

$$= \sqrt{\frac{\Sigma dx^2 - \frac{(dx)^2}{n_1} + \Sigma dx^2 - \frac{(dx2)^2}{n_2}}{n_1 + n_2 - 2}} \quad when \quad x - \bar{x} \neq 0$$

Illustration 28

A group of 5 patients treated with medicine A weigh 42, 39, 48, 60 and 41 Kgs ; second group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69 and 62 Kgs. Do you agree with the claim that medicine B increases the weight significantly.

Solution

A and B medicines have equal effect on the increase in weight.

x_1	$dx_1 = x_1 - 46$	dx_1^2	x_2	$dx_2 = X_2 - 57$	dx_2^2
42	- 4	16	38	- 19	361
39	- 7	49	42	- 15	225'
48	+ 2	4	56	- 1	1
60	+ 14	196	64	+ 7	49
41	- 5	25	68	= 11	121
			69	= 12	144
			62	+ 5	25
	0	290		0	926

$$\bar{x}1 = A + \frac{\Sigma dx_1}{n_1} = 46 + \frac{0}{5} = 46$$

$$\bar{x}2 = B + \frac{\Sigma dx_2}{n_2} = 57 + \frac{0}{7} = 57$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S} \times \sqrt{\frac{n_1\ n_2}{n_1 + n_2}}$$

$$S = \sqrt{\frac{\Sigma dx_1{}^2 \frac{\Sigma dx_1{}^2}{n_1} + \Sigma dx_2{}^2 - \frac{\left(\Sigma dx_2{}^2\right)}{n2}}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{290 + 926}{5 + 7 - 2}}$$

$$= 11.027$$

$$t = \frac{46 - 57}{11.026} \times \sqrt{\frac{5 \times 7}{5 + 7}}$$

$$= \frac{11.00}{11.027} \times \sqrt{\frac{5 \times 7}{5 + 7}}$$

$$= 1.70$$

Since the calculated value (1.70) is less than the table value 2.228 at to.05 for 10 d.f, the difference is insignificant. We conclude that the medicine A and medicine B do not differ significantly as regards their effect on increase in weight.

Illustration 29

The average number of articles produced by two machines per day are 200 and 250 with standard deviation 20 and 25 respectively on the basis of records of 25 days production.

Can you regard both the machines equally efficient at 0.01 level of significant?

Solution

Ho : No significant difference between the two machines in production.

$$\bar{x}_1 = 200 \quad S_1 = 20 \quad n_1 = 25$$
$$\bar{x}_2 = 250 \quad S_2 = 25 \quad n_2 = 25$$

S (pooled estimate of standard deviation on the basis of given standard deviations)

$$= \sqrt{\frac{n_1 S_1^{\,2} + n_2 S_2^{\,2}}{n_1 + n_2 - 2}} \quad (Where\ we\ are\ given\ n\ and\ S)$$

$$= \sqrt{\frac{25(20)^2 + 25(25)^2}{25 + 25 - 2}}$$

$$= 23.10$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S} \times \sqrt{\frac{n_1 \quad n_2}{n_1 + n_2}}$$

$$= \frac{200 - 250}{23.10} \times \sqrt{\frac{25 \times 25}{25 + 25}}$$

$$= \frac{-50}{23.10} \times 3.54$$

$$= -7.66$$

Since calculated |t| is more than table value (1.96 at 0.05 for 48 d.f) it, is highly significant and hence we reject H_o. We conclude that the performance of two machines differ significantly.

Two Dependent Samples Means

In the previous test, the two samples were independent. But there are many situations in which this condition does not hold true in the sense we have dependent

samples. Two samples are said to be dependent in nature, if the elements in one Sample are related to those in the other. The t-test for paired observations is definite as:

$$t = \frac{\bar{d} \times \sqrt{n}}{S}$$

$$S = \frac{\left(\Sigma d - \bar{d}\right)^2}{n-1}$$

$$or\ S = \sqrt{\frac{\Sigma d - \left(\bar{d}\right)^2 \times n}{n-1}}$$

Illustration 30

An IQ was administered to 5 persons before and after they were trained. The results are gi

Candidate	IQ before training	IQ after training
1	110	120
2	120	118
3	123	125
4	132	136
5	125	128

Test whether there is any change in IQ after the training programme.

Solution

Let H_o: Score before and after training is not the same. And let the score before and after training
denoted by x and y respectively.

X	Y	d (X - Y)	d^2
110	120	- 10	100
120	118	2	4
123	125	- 2	4
132	136	- 4	16
25	128	- 3	9
		- 17	133

$$\overset{\omega}{d} = \frac{\Sigma d}{n} = \frac{-17}{5} = -3.40$$

$$S = \sqrt{\frac{\Sigma d^2 - \left(\overset{\rho}{d}\right)^2 \times n}{n-1}}$$

$$= \sqrt{\frac{133 - (-3.40)^2 \times 5}{5-1}}$$

$$= \sqrt{18.8}$$

$$= 4.34$$

$$t = \frac{\overset{\omega}{d}\sqrt{n}}{S}$$

$$= \frac{3.40 \times \sqrt{5}}{4.34}$$

$$= -1.75$$

The calculated |t| *is* less than the table value of 2.776 at $t_{0.05}$ for 4 d.f. and hence accepted the hypothesis. We conclude that the students have benefited by the training programme.

Illustration 31

A group of 10 children were treated to find out how many digits they could repeat from after hearing them once. They were given practice for a week and were then tested. Is the difference between the performances of 10 children at the two tests significant.

Child	Test 1	Test 2
A	6	7
B	5	7
C	4	6
D	7	8
E	8	9
F	6	6
G	7	9
H	5	7
I	8	8

Solution

H_o: No significance difference between the performance of children with practice and those without practice.

Test 1	Test 2	d	d^2
6	7	- 1	1
5	7	-2	4
4	6	- 2	4
7	8	- 1	1
8	9	- 1	1
6	6	0	0
7	9	- 2	4
5	7	- 2	4
8	8	0	0
8	10	- 2	4
		- 13	23

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-13}{10} = -1.3$$

$$S = \sqrt{\frac{\Sigma d^2 - (\bar{d})^2 \times n}{n-1}}$$

$$= \sqrt{\frac{23 - (-1.3)^2}{10-1} \times 10}$$

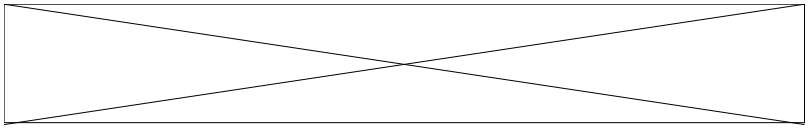
$$= 0.82$$

$$= \frac{\bar{d}\sqrt{n}}{S}$$

$$= \frac{-1.3 \times \sqrt{10}}{0.82}$$

$$= -5.01$$

The table value of 9 d.f at $t_{0.05}$ (2.262) is less than the calculated value $|t|$ of 5.01, and hence H_o is rejected, we conclude that the practice to children has better improvement.



Please use headphones

(b) F-Test

The main object of F-test is to discover whether the samples have come from the same universe. We can get the answer to this problem by the study of Analysis of Variance also called ANOVA, an obvious abbreviated word. For testing the difference of more than two samples, the F-test is an alternative to t-test which can be applied to test the difference of only two or less than two samples. The test consists of classifying and cross-classifying statistical results, and testing the significance of the difference between the samples statistics as well as among the samples statistics. The analysis of variance is studied by (a) one-way classification and (b) two-way classification.

If we observe the variation of the variables with one factor, it is known as one-day classification. If we observe the variation of the variables with two factors, it is called two-way classification. The estimate of population variance which is based on variation between the groups is known as the mean square between groups. The estimate of population variance that is based on the variation within group is known as the mean square within graphs. Since the F-test is based on the ratio of two variances, it is also known as Variance Ratio test. The variance ratio is denoted by F. It is given by:

$$F = \frac{\text{Mean Square between groups}}{\text{Mean Square within groups}}$$

If the calculated value of F is lesser than the table value, all the groups are drawn from a normal population. To calculate variance ratio, one must determine the following values :

(i) The total sum of squares

(ii) The sum of squares between groups

(iii) The sum of squares within groups

Illustration 32

To test the significance of possible variation in performance in a test between the programmer schools of a city; a common test was conducted to a number of students taken at random from the fifth class of each of the four schools concerned. The results are given below. Is there any significant in the means of samples?

A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Solution

Let us take the null hypothesis that there is no significant difference in the means of the samples

Samples

x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	x_4	x_4^2
8	64	12	144	18	324	13	169
10	100	11	121	12	144	9	81
12	144	9	81	16	256	12	144
8	64	14	196	6	36	16	256
7	49	4	16	8	64	15	225
45	421	50	558	60	824	65	875

(i) Correction factor = = $\frac{T^2}{N} = \frac{(220)^2}{20} = 2420$

(ii) Total sum of squares = Squares of all items - Correction factor

= 421 + 558 + 824 + 875 - 2420 = 258

(iii) Sum of squares between groups i.e. schools

$$= \frac{(\sum x_1)^2}{n} + \frac{(\sum x_2)^2}{n} + \frac{(\sum x_3)^2}{n} + \frac{(\sum x_4)^2}{n} - \frac{(T)^2}{N}$$

$$= \frac{(45)^2}{5} + \frac{(50)^2}{5} + \frac{(60)^2}{5} + \frac{(65)^2}{5} - 2420$$

= 405 + 500 + 720 + 845 - 2420

= 50

(iv) Sum of squares within groups i.e. Schools

= total sum of squares - sum of squares between groups

= 258 - 50 = 208

The above information may be presented in the form of a table, known as ANOVA table.

ANOVA Table

Sources of variation	Sum of squares	Degrees of freedom	Mean square	F
Between samples	50	k - 1 4-1 = 3	16.67	
Within samples	208	n - k 20 - 4 = 16	13.00	16.67 13.00
Total	258	19		

Thus, the calculated value of F is 1.28 and table value of F and at $\nu_1 = 3$ and $\nu_2 = 16$

0.05 is 3.24. The calculated value is less than the table value and hence accepts the hypothesis. W
conclude that the difference in the mean values of the samples (schools) have come from the sar
universe.

Illustration 33

A tea company appoints four salesmen A, B, C and D; and observes their sales in three seasons - Summer, Winter and Monsoon. The figures (in lakhs) are given as :

A tea company appoints four salesmen A, B, C and D; and observes their sales in three se
Summer, Winter and Monsoon. The figures (in lakhs) are given as :

Sales men

A	B	C	D
36	36	21	35
28	29	31	32
26	28	29	29

Is sales differ between the salesmen.

Solution

Let us take the hypothesis that there is no significant variation in sales between the salesmen.

Salesmen							
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	x_4	x_4^2
36	1296	36	1296	21	441	35	1225
28	784	29	841	31	961	32	1024
26	676	28	784	29	841	29	841
90	2756	93	2921	81	2243	96	3090

Correction factor = $\frac{T^2}{N} = \frac{360^2}{12} = 10800$

Total = Squares of all items –correction factor

= 2756 + 2921+ 2243 + 3090 - 10800

= 210

Sum of squares between salesmen = Divide the squares of total by n, add all such figures — correction factor

$$= \frac{(90)^2}{3} + \frac{(93)^2}{3} + \frac{(81)^2}{3} + \frac{(96)^2}{3} - 10800$$

= 2700 + 2883 + 2187 + 3072 - 10800 = 42

Within salesmen = Total sum of squares- Sum of squares between salesm

= 210 - 42 = 168

ANOVA Table

Variation	SS	DF	MS	F - ratio
Between salesmen	42	3	14.00	$\frac{14}{21}$
Within salesmen	168	8	21.00	= 0.67
	210	11		

For v_1= 3 and v_2 = 8 - at 0.05 - 4.07

Conclusion

The calculated value of F between salesmen is 0.67 which is less than the table value, hence conclude that the sales of different salesmen do not differ significantly.

- End Of Chapter -

LESSON - 12

CHI SQUARE TEST

Chi Square Test

The significance for small samples has been tested by t and F based on the assumption that the samples were drawn from normally distributed population. Since testing the significance requires an assumption about the parameters (i.e. population values such as mean, standard deviation, correlation etc.) hence t and F tests are called PARAMETRIC tests. In reality, all distributions of variables pertaining to social, economic and business situations may not be normal. The limitation of t and F tests has led to the development of a group of alternative techniques (tests) known as NON-PARAMETRIC or distribution - free methods.

In the study of non-parametric tests, no assumption about the parameters of the population from which we draw samples is required. An increasing use of non-parametric tests in economic and business is on account of the three reasons namely:

(i) the non-parametric test are distribution-free,

(ii) they are computationally easier to handle and understand than parametric tests ; and

(iii) they can use with types of measurements that prohibit the use of parametric tests.

Of course, the non-parametric tests are popular in application but not superior to the parametric methods. In fact, in situations where both tests apply, the non-parametric tests are more desirable than the parametric tests.

The square of a standard normal variable is called a Chi-square (pronounced as ki-square) variety with one degree of freedom. The chi-square test is used in a very large number of cases to test the accordance between fact and theory. In other words, it describes the magnitude of discrepancy between theory and observation. The chi-square has three applications namely:

(i) test of goodness of fit,

(ii) test for independence of attributes and

(iii) test for population variance.

a) Test of Goodness of Fit:

A good compatibility between theory and experiment proved by the statistic chi-square test which is defined as

$$X^2 = \sum_{i=1}^{n} \frac{(Oi - Ei)}{Ei}^2$$

where Oi refers to observed frequencies and

Ei refers to expected frequencies.

Computation and Observation

The value of chi-square is derived by:

(i) taking the difference between an observed frequency and expected frequency,

(ii) squaring this difference,

(iii) dividing the squared difference by expected frequency,

(iv) add the values so obtained to compute chi-square.

Then compare this value to the table value with desired degrees of freedom. The degrees of freedom we mean the number of classes to which the values can be assigned arbitrarily. The degrees of freedom are denoted by d.f. at the level of n-1. In a contingency table, the d.f. are (r-1) (c-1) where r and c refers to the number of rows and columns respectively. The expected frequency for any cell is E = RT x CT/N.

If the calculated value of chi-square is less than the table value of chi-square, then it is said to be non-significant at the required level of significance. This implies that the discrepancy between observed (experiment) values and expected (theory) values may be attributed to chance i.e. variations of sampling. In other words, data do not provide any evidence against the hypothesis which may be accepted. In brief, we may conclude that there is good correspondence or good fit between theory and experiment. On the other, if the calculated value is greater than the tabulated value, it is said to be significant. In other words, we conclude that the discrepancy between observed and expected frequencies cannot be attributed to chance, hence the experiment does not support the theory.

Conditions for applying test

i. The total frequency should be sufficiently large, say greater than 50.

ii. The sample observations should be independent in the sense no individual item should be included twice or more in the same sample.

iii. The constraints must be linear. Equations containing no squares or high powers of the frequencies are linear constraints (such as S0 =S*E).*

iv. Expected frequency should be 10, but not less than 5. If it is less than 5, the technique of pooling is to be applied.

Illustration 34

The number of automobile accidents per month was as follows: 48, 32, 80, 10, 54, 42, 60, 26, 38, 20. Are these frequencies in agreement with the belief that accident conditions were same during 10 months period.

Solution

$$x^2 = \sum_{n}^{i-1} \frac{(o_i - E_i)^2}{E_i}$$

we can find the chi-square value

$$x^2 = \frac{(48-41)^2}{41} + \frac{(32-41)^2}{41} + \frac{(80-41)^2}{41} + \frac{(10-41)^2}{41} + \frac{(54-41)^2}{41} + \frac{(42-41)^2}{41} + \frac{(60-41)^2}{41} + \frac{(26-41)^2}{41} + \frac{(38-41)^2}{41} + \frac{(20-41)^2}{41}$$

= 1.1951 + 1.9756 + 27.0976 + 23.4390 + 4.1219 + 0.0024

+ 8.8049 + 5.4878 + 0.2195 + 10.7561 = 93.0999 = 93.100

Since the calculated value is more than the table value of 16.919 for 9 d.f. at 0.05, it is significant and null hypothesis is rejected. We conclude that the accidents are certainly not uniform during the period of 10 months.

Illustration 35

A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in second class and 20 got a first class. Are these figures commensurate with the general examination result policy which is in the ratio of 4:3:2:1 for the various categories respectively.

Solution

HO:

The observed figures do not differ significantly to the ratio of 4 : 3 : 2 : 1. Frequency distribution of the result of 500 students is as follows:

Category	Observed frequency (O)	Expected frequency (E)	$(O - E)^2 E$
Failed	220	$\frac{500 \times 4}{10} = 200$	2.0000
III Class	170	$\frac{500 \times 5}{10} = 150$	2.6667
II Class	90	$\frac{500 \times 2}{10} = 100$	1.0000
I Class	20	$\frac{500 \times 1}{10} = 50$	18.0000
Total	500	500	23.6667

Since the calculated value of chi-square (23.667) is more than the table value of 7.815 for 3 d.f. at 0.05, the difference is significant and Ho is rejected. We conclude that the data are not commensurate with the general examination result policy.

Illustration 36

Records show the number of male and female births in 800 families having four children are given below.

Number of births:	Male	0	1	2	3	4
	Female	4	3	2	1	0
	Frequency	32	178	290	236	64

Test whether the data are consistent with the hypothesis that the binomial law holds and the chance of male birth is equal to that of a female birth.

Solution

H_o: The data are consistent with the binomial law of equal probability for male and female births and we have $p = q = 0.5$ (p denotes the probability of a male birth). If this is true, the chances of males and females according to binomial probability law are given. $\left(\frac{1}{2}+\frac{1}{2}\right)^4$

How to fit binomial distribution: Suppose a random experiment consists of n trials, satisfying the conditions of binomial distribution and suppose this experiment is repeated N times, then the frequency of r successes is given by the formula.

$P(r) = N\,p(r) = N \times n_{cr}\, p^r\, q^{n-r}$

$r = 0, 1,2,3....n$

$= 800 \times 4_{cr}\,(1/2)^r\,(1/2)^{4-r}$

$= 800 \times 4_{cr} \times (1/2)^4$

$= 50 \times 4_{cr}$, $(r = 0,1,2,3,4)$

By substituting r = 0,1,2,3,4 successively we get the Binomial frequencies as given below:

Male birth (r)	Expected frequency		
0	$50 \times 4_{c0}$ $\left(50 \times \frac{4}{4}\right)$	50 x 1	50
1	$50 \times 4_{c1}$ $50 \times \frac{4}{1}$	50 x 4	200
2	$50 \times 4_{c2}$ $50 \times \frac{4 \times 3}{2 \times 1}$	50 x 6	300
3	$50 \times 4_{c3}$ $50 \times \frac{4 \times 3 \times 2}{2 \times 1 \times 1}$	50 x 4	200
4	$50 \times 4_{c4}$ $50 \times \frac{4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1}$	50 x 1	50

Now, calculate goodness of fit.

Male births	O	E	$\frac{(O-E)^2}{E}$
0	32	50	6.48
1	178	200	2.42
2	290	300	0.33
3	236	200	6.48
4	64	50	3.92
	800	800	19.63

Since the calculated value of chi-square is 19.63 which is greater than the table of 9.488 for 4 d.f at 0.05, it is significant. Thus, the difference between the observed and expected frequencies is significant, hence null hypothesis is rejected. We con-

clude that the equal male and female births is wrong and the binomial distribution with p = q = 0.5 is not a good fit to the given data.

Illustration 37

A set of 5 coins is tossed 3200 times, and the number of heads appearing each time is noted. The results are given below.

No. of heads :	0	1	2	3	4	5
Frequency :	80	570	1100	900	500	50

Test the hypothesis that the coins are unbiased.

Solution

Solution

Ho: Coins are unbiased. If this is true the expected frequency according to binomial probability law is $\left(\frac{1}{2}\times\frac{1}{2}\right)^5$

$P(r) = N\ p\ (r) = N\times {}^{n}c_{r} p^{r} q^{n-r}$

$r = 0,1,2,3,4,5$

$= 3200\ {}^{5}c_{4} \left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{5-r}$

$= 3200 \times {}^{5}c_{r} \left(\frac{1}{2}\right)^{5}$

$= 100\times {}^{5}c_{r}$; $r = 0,1,2,3,4,5$

No. of heads	Expected frequency			
0	$100 \times 5c_0$	=	100 x 1	100
1	$100 \times 5_{C1}$	=	100 x 5	500
2	$100 \times 5_{c2}$	=	100 x 10	1000
3	$100 \times 5_{C3}$	=	100 x 10	1000
4	$100 \times 5_{C4}$	=	100 x 5	500
5	$100 \times 5_{c5}$	=	100 x 1	100
Testing of Goodness of fit				
0	E		$(O - E)^2 E$	
80 570 1100 900 500	100 500 1000 1000 500		4.00 9.80 10.00 10.00 0	
50	100		25.00	
3200	3200		58.80	

$x^2 = 58.80$

The calculated value is more than the table value of 11.070 for 5 d.f at 0.05. Hence the hypothesis is rejected. We conclude that coins are biased.

Illustration 38

In experiment of pea-breeding, Reddy got the following frequencies of seeds: 315 round and yellow, 101 wrinkled and yellow, 108 round and green, 32 wrinkled and green, total 556. Theory predicts that the frequencies should be in the proportion of '9:3:3:1. Examine the correspondence between theory and experiment.

Solution

Ho:

Theory (expectation) and experiment (observation) corresponds each other.

Colour	Experiment frequencies (O)	Theory frequencies (E)	$\frac{(O-E)^2}{E}$
Round and yellow	315	312.75	0.0162
Wrinkled and yellow	101	104.25	0.1013
Round and green	108	104.25	0.1349
Wrinkled and green	32	34.75	0.2176
	556	556.00	0.4700

For 3 d.f at 0.05 = 7.81

Result:

Ho is accepted. A good correspondence between experiment and theory may be observed.

(b) Test of Independence of Attributes

An information is said to be attribute when it is qualitative/The chi-square helps to find out whether two or more attributes are associated or not. Examples of attributes are: beauty, colour, failure etc. The expected frequency for an attribute for each of the cell in contingency table is determined as:

$$\text{Expected frequency of a cell} = \frac{\text{corresponding cell row total} \times \text{colum total}}{n}$$

Contingency Table

A table having R rows and C columns is known as contingency table. Each row corresponds to a level of one variable, each column to a level of another variable. Entries in the body are frequencies with which each variable combination occur. A table having 3 rows and 2 columns is called a 3 x % contingency table while 2 rows and 3 columns is known as a 2 x 3 contingency table. Thus, the table depends on the number of rows and columns.

Illustration 39

Out of a sample of 120 persons in village, 76 persons were administered a new drug for preventive influenza and out of them, 24 were attacked by influenza Out of those who were not administered the new drug, 12 persons were not affected by influenza. Prepare (i) 2 x 2 table showing actual and expected frequencies, (ii) use chi-square test for finding out whether the new drug is effective or not.

Solution

Ho: New drug is effective in controlling influenza,

(i) Preparation of 2 x 2 table

	Attacked	Not-attacked	Total
Administered Non-administered	24 (35.47) 32 (20.53)	52 (40.53) 12 (23.47	76 44
Total	56	64	120

Expected frequency for 24 $= \frac{RTXCT}{N} = \frac{76 \times 56}{120} = 35.47$

$$X^2 = \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E}$$

$$= \frac{(24-35.47)^2}{35.47} + \frac{(32-20.53)^2}{20.53} + \frac{(52-40.53)^2}{40.53} + \frac{(12-23.47)^2}{23.47}$$

= 3.7090 + 6.4082 + 3.2680 + 5.6053

= 18.9687

Since the calculated value is greater than the table value of 3.84 for 1 d.f at 0.05, hence the null hypothesis is accepted. We conclude that the new drug is definitely effective in controlling i.e. preventing the disease (influenza).

The computation of chi-square as above is quite tedious and consuming process. The chi-square can conveniently compute by using the alternative formula.

$$X^2 = \frac{N\ (ad-bc)^2}{(a+b)(b+d)(a+b)(c+d)}$$

By substituting the values in formula we get,

$$X^2 = \frac{120((24\times12)-(52\times32))^2}{56\times64\times75\times44}$$

$$= \frac{120\ (1376)^2}{56\times64\times75\times44}$$

= Antilog 120 + 2 log 1376 - log 56+ log 64 + log 76+ log 44

= Antilog 8.3564 – 7.0786

= 18.96

The chi-square value of 18.96 is more approximately the same as obtained earlier,

(c) Test for Population Variance

The chi-square test is also used to know whether a random sample has been drawn from a normal population with mean and variance. The statistic

$$X^2 = \frac{\sum (x-\bar{x})^2}{\theta^2} = \frac{n s^2}{\theta^2}$$

which follows chi-square distribution with n-1 d.f. If the sample size is large, say greater than 30, we can use Fisher's approximation namely,

$$Z = \sqrt{2X^2} - \sqrt{(2n-1)}$$

Illustration 40

Weights, in kgs, of 10 students are given below:

38, 40, 45, 53,47,43,55,48, 52,49. Can we say that variance of the distribution of weights of all students from which the above sample of 10 students was drawn, is equal to 20 square kgs?

Solution

Set up the null hypothesis, Ho : $\theta^2 = 20$.

Weight in kgs	$(x-\bar{x})$ i.e., $(x-47)$	$(x-\bar{x})^2$	
38	-9	81	
40	-7	49	
45	-2	4	
53	+6	36	
47	0	0	
43	-4	16	
55	+8	64	
48	+1	1	
52	+5	25	
49	+2	4	
470	0	280	

Now , we have

$$n = 10,\ \sum (x - \bar{x})^2 = 280 \text{ and } \bar{x} = \frac{470}{10} = 47$$

$$X^2 = \frac{\sum (x - \bar{x})^2}{\theta^2} = \frac{280}{20} = 14.00$$

Chi-Square Table value for 9 d.f at 0.05 = 16.92.

The calculated value is less than the table value, it is not significant and hence accept the null hypothesis. The given data is consistent with the hypothesis that the variance of the distribution of weights of 10 students in the population is 20 square kgs.

Illustration 41

A random sample of size 20 from a population gives the sample standard deviation of 6. Test the hypothesis that the population standard deviation is 9.

Solution

Set up H_o that the population standard deviation is 9. We are given,

n = 20 and s = 6

$$X^2 = \frac{n s^2}{\theta^2} = \frac{20 \times (6)^2}{(9)^2} = \frac{20 \times 36}{81}$$

$$= \frac{720}{81} = 8.88$$

The calculated value is less than the table value of 30.144 for 19 d.f at 0.05, hence H_o may he accepted. In other, words, the population standard deviation, 9 may be accepted.

Illustration 42

Test the hypothesis that variance = 64 given that sample variance 100 for a random sample of size 51.

Solution

Ho: Population variance is 64. We are given

n = 51 Variance (population = 64)

Variance (sample) = 100

$$X^2 = \frac{n s^2}{\sigma^2} = \frac{51 \times 100}{64} = \frac{5100}{64} = 79.69$$

Since the sample size is large i.e. more than 30, Fisher's approximation may be used to chi-square distribution.

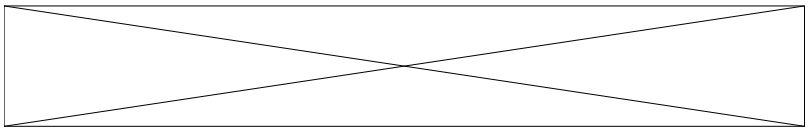
$$Z = \sqrt{2X^2} - \sqrt{2n-1}$$

$$= \sqrt{2 \times 76.69} - \sqrt{2 \times 51 - 1}$$

$$= 12.62 - 10.05 = 2.57$$

Z table value = 1.96 at 0.05 level of significance.

Since Z is greater than 1.96, hence the null hypothesis is rejected. In other words, that the random sample cannot be regarded as drawn from the population with variance 64.



Please use headphones

EXERCISES

1. Explain the importance of Theory of Estimation in decision - making.
2. What do you mean by Point Estimation and Interval Estimation? Explain with illustrations.
3. What is standard error? State the application of standard error in large sample test of statistical hypothesis.
4. Describe the important properties of a good estimator.
5. What is meant by the sampling distribution of a statistic? Describe briefly the sampling distribution of mean and variance.
6. What do you understand by small sample tests? How are they different from large sample tests?
7. What do you understand by t-test, F-ratio and Chi-square. Discuss their application.

8. A sample of 100 families selected at random gives an average income of Rs.8000 with a standard deviation of Rs. 3000. Estimate the confidence interval of mean income of families at 95 per cent and 99 per cent.Ans: at 95 % — Rs. 7412 and Rs. 8588 at 99%,- Rs. 7226,andRs. 8774
9. 50 out of 500 machine parts tested are found to be defective. Estimate the 95per cent confidence interval for the proportion of defective machine parts.

   Ans: 7.50 percent and 12.50 per cent.

REFERENCE BOOKS

Agarwal, B.L. "Basic statistics", Wiley Eastern Limited, New Delhi, 1994.

Gupta, S.P. "Statistical Methods", Sultan Chand and Sons, New Delhi, 1992.

Lehman, E.L."Testing Statistical Hypothesis", Wiley Eastern Limited, New Delhi, 1976.

Levin, R.I. "Statistics for Management", Prentice - Hall of India, New Delhi, 1987

Reddy, C.R. "Quantitative Methods for Management Decision?, Himalaya Publishing House, Bombay, 1990.

Rao, C.R. "Linear Statistical Inference and its Applications", Wiley Eastern Limited, New Delhi, 1977.

- End Of Chapter -

LESSON - 13

REGRESSION AND REGRESSION MODEL

Introduction

The study of a single variable with regard, to its distribution, moments etc. is one facet of statistical studies. Another aspect of statistical analysis is the joint study off two or more variables in "respect of their interdependence or functional relationships while studying the interdependence, we study the correlation between two or more variables. The statistics does-not explain the causative-effect between the variables but it is being explored through other considerations. The theory of correlation and association between variables was largely developed by Karl Pearson and G.Undy Yule in early twentieth century. The idea of dependence between two or more quantitative normal variants further lead to the theory of regression. In regression analysis, one has to deal with two or more interdependent Variables. Among the variables under consideration, we have two sets of variables, one is known as dependent variable and the other independent variable(s).

For instance weight of a person is related to his height. So we may find the regression of height on weight or weight on height. Age of wife and age of husband at the time of marriage is another example of interdependent variables. But such situations seldom prevail. Mostly, there is a variable which depends on one or more variables. For instance, the yield of a crop depends on fertilizer dose. In this situations yield is the dependent variable and fertilizer dose is an independent variable, The production cost of an unit depends on the cost of raw-material, labour cost, cost of electricity, transportation etc. In this example, production cost is the dependent variable and the cost, of raw material, labour cost, cost of electricity and transportation cost etc. are independent variables.

From the above discussion it is apparent that in regression analysis we deal with two types of variable, the one dependent variable and the other independent variable. Not going beyond the scope of curriculum, it would not be wrong to say that; there is one dependent variable, called response variable, which depends on one or more variables, so called independent variable(s). The independent variables are also termed as regressors, predictors, explanatory variables etc.

The concept of regression was first developed by Francis Galton in a study of inheritance of structure in human being. To prove this biometrical fact Karl Pearson found the regression of son's height on father's height. But soon the use of regression technique was extended in large number of sciences like Economics, Sociology, Psychology, Medical Sciences, Zoology, Breeding, Agronomy, Management etc. Now adays it is one of the most frequently used tool of statistics.

Objective

Literal meaning of the word regression is to progress or to step back. Sir Francis Galton used this word towards mediocrity in hereditary statue. But now a days it is used much wider without giving any heed towards its old conceptual sense. The objectives of regression can be delineated in the following manner.

1. The foremost objective of the regression is to establish the actual functional relationship between the dependent and independent variables.
2. To estimate the value of the dependent variable for a given value of X.
3. Regression equation can very well be used as prediction equation. A value of the dependent variable can easily be estimated for the given value(s) of the independent variable or variables.
4. It helps to find the trend in analysis of time series.
5. Regression is aimed for projections like population projection, production of cereal crops etc, some of the writers are quoted here which throw light on the objectives of regression.

Werner Z. Hissch (Introduction to modern statistics)

Regression analysis measures the nature and extent of this relation, thus enabling us to make predictions.

Wallis and Reberts (Statistics: A new approach)

It is often more important to find out what the relation actually is, in order to estimate or predict one variable (the dependent variable); and the statistical technique appropriate to such a case is called regression analysis.

J.R. Stockton (Introduction to Business Economics and Statistics)

The device used for estimating the value of one variable from the value, of the other consists of a line through the points drawn in such a manner as to represent the average relationship between the two variables. Such a line is called the line of regression. Once the objective of regression analysis is clear, the problem confronted by one is to establish a suitable and appropriate statistical model which has to be fitted in by the use of actual data.

Regression Model

There is a difference between a mathematical equation and statistical function. Mathematical equation simply describes the intrinsic relationship between two or more unknown variables. Whereas the statistical function considers the two type of variables, involves parameters and an unknown error term whose distribution is always taken into consideration.

A statistical model for regression equation between a dependent variable Y and the independent variables X_1, X_2, ... X_k involving the parameter 0_0, 0_1, 0_2.... 0_k and the error terms can in general, be given as:

$Y= [X_l, X2,....,X_k / 0_0,0_1,0_2,.....0_k] + e$ (1)

Where Y indicates the form of the function. It may be linear or curvilinear, that too in a specific form, is a random error which is usually taken to be distributed with mean zero and variance σ_ε^2, notationally $N(0, \sigma_\varepsilon^2)$, Equation (1) is called the regression equation of Y on X, which means Y is the dependent variable and X's are the independent variables. The necessity of introducing the error term arises due to the fact that the parameter of the regression equation cannot be estimated without it since the observed value of the dependent variable will rarely agree with the expected values. How so ever exactly the independent variable X_1, X_2,.... X_k might have been measured. The statistical model without the error terms, that is $Y = (X_1, X2,.... X_k / 0_0,0_1,0_2,.0_k....)$ as a special case is known as a deterministic model.

If the relationship between Y and X's is linear, it is known as a linear regression equation. In this equation none of the variables of Y and X's has power more than 1. If it has a term with power other than 1, it represents a-curve and is called curvilinear regression. For instance an equation between Y, X_1 and X_2-of the type represents a curvilinear regression for instance an equation between Y,X_1 and X_2 of the type.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Is a linear regression equation and of the type

$$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 X_1 X_2 + \beta_3 X_2^2 + \varepsilon$$

Represents a curvilinear regression.

Assumptions about the regression model

There are four assumptions which are usually made in a regression model. For simplicity we are considering only two variables namely, the dependent variable Y and an independent variable X.

i. For each selected X, Y's are distributed normally and independently with mean $\mu_{y/x}$ and variance $\sigma^2_{y/x}$. If we consider the simple line are equation between Y and X, then $\mu_{y/x} = \beta_0 + \beta_1 x$ and $\sigma^2_{y/x} = \sigma^2_y x$

ii. The population of values of Y corresponding to each selected X has mean which lies on line,

$$\mu = \beta_0 + \beta_1 (x - \bar{x})$$

$$= \beta_0 + \beta_1 x \text{ where } X - \bar{X} = x$$

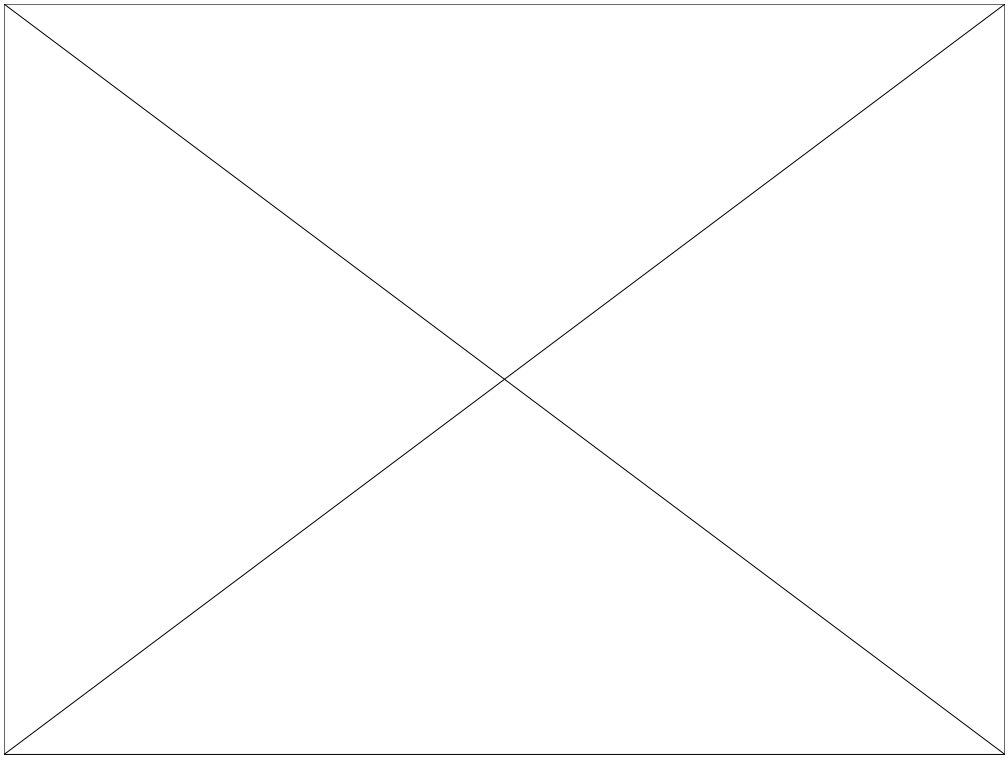
iii. There is no error in the measurement of the variable X and is known exactly.

iv. Variances $\sigma^2_{y/x}$ are homogeneous in linear regression model and is same as σ^2_ε

Selection of a regression model

There is no thumb rule which can be given to select a regression model. It is the nature of variables, kind of observation, purpose of study, long experience etc. on the-basis of which, a regression model has to be selected. More frequently used model is a linear model. Hence, most of our discussion will be centered on linear model. In case of two variables, there is one dependent variable Y and one independent variable X, a good idea of the relationship between Y and X can be had from a scatter diagram. A simple linear regression model with Y as dependent variable and X as independent variable is given as

$$Y = \alpha + \beta X + \varepsilon \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$



Please use headphones

This known as the simple regression line of Y on X.

In case of linear regression with two Independent variables X_1 and X_2, the regression model is given as.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \times X_2 + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

At the same time, a most popular curvilinear regression is a second degree equation given as

$$Y = \alpha + \beta X + \gamma X^2 + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

This equation represents a parabola. ,

In the same way any other linear or curvilinear regression model can be given.

Scatter diagram

When we wish to set up a regression equation between two variables Y and X, the idea about the type of relationship between the two can easily be obtained by plotting the n paired values (X_1, Y_1), (X_2, Y_2), (X_n, Y_n) on a graph paper. If these points lie in a straight line or lie in the close vicinity of this line, then a linear relationship is considered appropriate. If some other known pattern is observed the regression equation has to be chosen accordingly. While plotting the points (Xi, Yi.) for = 1,2, n,

one should always take Y, the dependent variable along ordinate (Y-axis) and the independent variable X along abscissa (X-axis) by choosing suitable scales.

If the plotted points do not show any pattern, then it is considered that4he two variables are independent. In this situation rarely more than 2,3 points lie in a straight line or in the pattern of a known curve.

Now we give below two scatter diagrams and discuss then in brief.

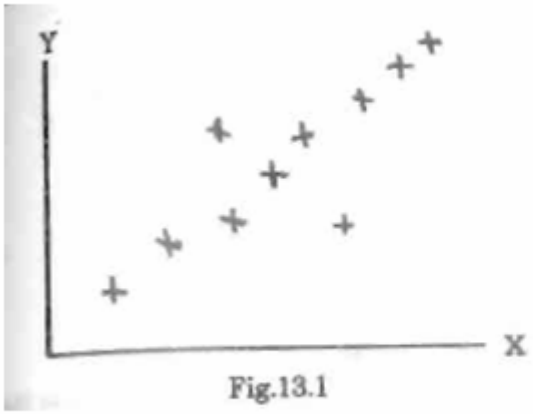


Fig.13.1

Scatter diagram showing linear trend

Fig.13.2

Scatter diagram showing no pattern

In fig. 13.1 few points lie on the straight line and the other are just above or below the line. A line of best fit will be one for which the perpendicular distance of all the points from this line is minimum. If the distance from above to the line as positive and from below the line as negative; the sum of the distances is almost zero.

In fig.13.2, it is easy to note that hardly any three points lie in a straight line or show any known pattern. Hence, .the variables X and Y are treated as independent and no regression equation is possible. In such a situation, no path is discernable.

Example 1:

Following are the scores out of 100 obtained in a test by the sales representatives and their sales performance in lac rupees.

Scores:	40	45	65	55	70	85	35	60	75	80
Sales :	12	14	22	18	31	34	15	20	24	30

To know the kind of relationship, we plot the points on graph paper, the graph is as follows:

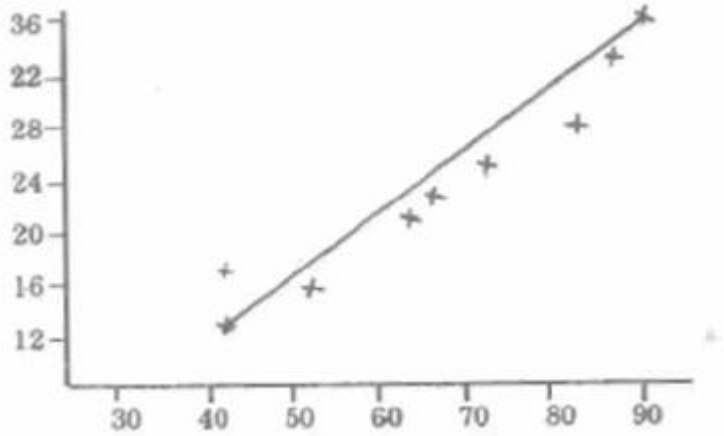


Fig.13.3

The figure reveals that the points lie on and around a straight line. Hence, simple linear regression line can be fitted to the data.

Example 2: The following data give the years of service and ratings of employees.

Employee : A B C D E F G H I J K L M N O P Q R

Year of

Service (X) : 1 7 9 5 4 3 2 5 6 8 11 12 10 7 6 5 3 14

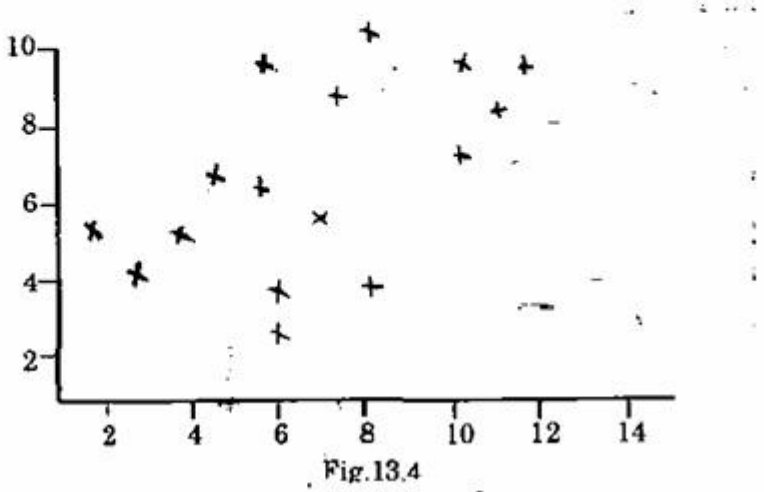


Fig.13.4

The paired observation (x, y) plotted on the graph paper clearly indicate that there is no set pattern shown by the points and hence no equation can be chosen to fit *m* the data.

QUESTIONS

1. Write a brief note on the importance of regression

2. Discuss a statistical regression model.

3. What are the assumptions made in a regression model?

4. What do you understand by independent and dependent variables?

5. Why do we call a regression equation, a predictions equation?

6. Explain scatter diagram and its utility.

7. What is meant by simple linear regression line?

8. What are the objectives of regression analysis?

9. Quote the statements of any two well known statisticians.

10. The heights (in cms) and weights (in kgs) of a random sample of twelve adult males are given below:

Height(X):	177	163	173	182	171	168	174	155	184	170	172	168
Weight(Y):	71	67	77	85	69	62	73	56	65	72	65	68

Draw a scatter diagram and comment on it.

11. The following figures relate to the number of units produced by workers and their ages.

Units Produced (Y) :	26	37	40	35	30	30	40	27	30	35	45
Age in yrs (X):	19	12	24	24	18	20	28	20	15	30	40

Plot the data and prepare a scatter diagram.

What do you infer from this scatter diagram?

12. What are other names given to the dependent and independent variables?

- End Of Chapter -

LESSON - 14

FITTING OF SIMPLE REGRESSION EQUATION

Simple regression line

A linear equation between two variables X and Y represents a straight line. A simple regression--line involves two variables Y and X. If Y is the dependent (response) variable and X is the independent variable, the simple regression line is given by the equation.

$$Y = \alpha + \beta X + \varepsilon \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

The choice of the dependent and independent variables depends on the nature of the variables: for example, age of persons and ft. Q. are-two variables. Out of these two, age is the independent, variable because it goes on increasing with the passage of time and does not depend on any factor and I.Q. increase with age as a person learns many things through education and experience. Hence, in this case age has to be taken as the independent variable X and I. Q. the dependent variable Y.

The constants α and β in equation (1) are called the parameters of the regression equation. Where ε is a .random variable, known as error. This error term is introduced, due to the fact that, the-actual value of Y rarely equals to the estimated value of Y. Model (l) is known, as probabilistic model as ε is distributed normally with mean zero variance σ_ε^2 i.e, ε N $(0,\sigma_\varepsilon^2)$. A similar-distribution is followed by Y i.e. $Y \quad N(0, \sigma_y^2)$

Equation (1) is similar to the intrinsic equation of straight line" in coordinate geometry,

Y= mx + c. Since the properties of a line do not change, we can say by symmetry that in equation *(1)* α is the intercept which the regression line cuts from the Y-axis and β is the slope of the line. Equation (1) is the regression line of Y on X and β is known as the **regression coefficient** of Y on X and is usually denoted β_{yx}. If the suffixes are not given, they are to be understood. But when it is to be specified that it is the regression Y on X, the suffixes are to be indicated.

Definition of regression coefficient

The regression coefficient is a measure of change in dependent variable Y corresponding to an unit increase in

independent variable. X.

Explanation: Suppose the value of $\beta_{yx} = 2.34$. It means that if X is increased by unit, Y is increased by 2.34. On the contrary if $\beta_{yx} = -3.5$ it means that if X is increased by unit, Y is decreased by 3.5 on an average;

Two regression equations

It is not necessary that we are always .to find out the regression of Y on X; There are many situations when the variables Y and X are such that either of the two variables can be taken as dependent variable, and the other as independent variable. For example, height and weight of persons are two variables in which height depends on weight and weight depends on height. So in such a situation we can find U the regression of height on weight or weight on height. In this way, we get two regression equation, the one of height (Y) on weight (x) or the other of weight (X) on height (Y). A word of caution is important at this juncture. A regression of Y on

X, $Y = \alpha + \beta_{\chi}$ written as $\chi = \frac{(Y - \alpha)}{\beta}$ does not become regression equation of X on Y. It remains

the regression of Y on X All the more an equation cannot represent both the equations as the mode of dependence of Y on X is not the same as the mode of dependence of X on Y. the regression equation of X on Y can be given as,

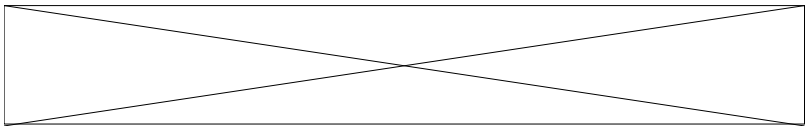
$$\chi = \alpha_1 + \beta_1 Y + \varepsilon_1 \qquad (2)$$

In the above equations as is the intercept which the line cuts from the X - axis and β_1 is the slope of the line from the Y - axis. β_1 is usually denoted β_{xy} and is known as the **regression coefficient of X and Y.**

Definition of β_{xy} : It is a measure of an average change in dependent variable X corresponding to an unit increase in independent variable Y.

Fitting of a regression line

$$Y = \alpha + \beta \chi + \varepsilon \qquad (3)$$



Please use headphones

Fitting of the regression equation means the estimation of the parameters a and through b the paired observations (x, y). By paired observation we mean the observations taken on He same item or individual or the already paired items. For instance, height and weight of the same person are to be taken in pair; income and expenditure of the same person are to form a paired observation etc. The best estimates; of a and b will be those which minimise the error. As we all know, the least error is zero. But such an ideal situation is hard to achieve. Hence, effort is made to estimate a and b such that the error is minimum. There are many methods of estimation, but we shall discuss here only the least square method of fitting the

regression equation. Least square method of estimation was given by the mathematician Legendre.

Let there be n pairs of observations, (X_I, X_I) $(X_2, X2)$ $(X_3, X3)$..... (X_n, Y_n). From these n pairs, we will estimate the values of a and b so that the error is minimum.

From equation (3)

WE GET, $\varepsilon = Y - \alpha - \beta \chi$ .....(4)

For the i^{th} pair of observation (X_i, Y_i)

$$\varepsilon_i = Y_i - \alpha - \beta \chi_i \qquad \text{.....(5)}$$

ε_i, will be positive if Yi is greater than $(\alpha + \beta \chi_i)$ and negative if Y_i is less than $(\alpha + \beta \chi_i)$ avoid the intricacy of the sign of error term, we square both sides of the equation and then take the sum over all pairs of values.

Thus,

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta \chi_i)^2 \qquad \text{.....(6)}$$

In this way we are left with the magnitude of the error term only. To minimise by the; method of least square, we differentiate equation (6) partially with respect to α and β respectively and equate then to zero. Let us suppose $\sum \varepsilon_i^2 = Q$ also replace α and β estimated values a and b. thus.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \alpha\beta\chi_i)^2$$

$$\frac{\partial Q}{\partial \alpha} = -2\sum_{i=1}^{n} (Y_i - \alpha - bx_i) = 0 \qquad \text{.....(7)}$$

$$\frac{\partial Q}{\partial \beta} = -2\sum_{i=1}^{n} (Y_i - \alpha - b\chi_i)(-\chi_i) = 0 \qquad \text{.....(8)}$$

From equation (7)

$$-2 \neq 0,$$

$$\therefore \sum_{i=1}^{n} (Y_i - \alpha - bx_i) = 0$$

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} a + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} x_i^2 \qquad \ldots\ldots\ldots(9)$$

From equation (8),

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \qquad \ldots\ldots\ldots(10)$$

Equations (9) and (10) are called as the normal equations. Solving (9) and (10) we get the values of 'a' and 'b' in terms of known observations.

In equation (9) divide both sides by n.

This gives,

$$\frac{1}{n}\sum_{i=1}^{n} y_i = a + b\frac{1}{n}\sum_{i=1}^{n} x_i \qquad \ldots\ldots\ldots(11)$$

or $\bar{y} = a + b\bar{x}$

or $a = (\bar{y} - b\bar{x})$ $\ldots\ldots\ldots(12)$

Now substituting the value of 'a' from equation (11) in equation (10), we get

$$\sum_{i=1}^{n} x_i y_i = \left(\frac{1}{n}\sum_{i=1}^{n}, x_i - b\frac{1}{n}\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - b\frac{1}{n}\sum_{i=1}^{n} x_i \left(\sum_{i=1}^{n} x_i\right)^2 + b\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - b\frac{1}{n}\sum_{i=1}^{n} x_i \left(\sum_{i=1}^{n} x_i\right)^2 + b\sum_{i=1}^{n} x_i^2$$

Therefore,

$$b = \frac{\sum_{i=1}^{n} x_i y_i = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \quad \ldots\ldots\ldots(13)$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(yi - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \ldots\ldots\ldots(14)$$

Dividing numerator and denominator of (14) by n we obtain,

$$b = \frac{\frac{1}{n} - \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{Cov(x, y)}{Var(x)} \quad \ldots\ldots\ldots(15)$$

$$= \frac{\sigma_{xy}}{\sigma_x^2} \quad \ldots\ldots\ldots(16)$$

Where σ_{xy} is the covariance between x and y and σ_x^2 is the variance of x. No matter, in which form the formula for b is used, the value of the regression coefficient b remains the same.

Of course the choice of the form of the formula depends on the convenience of calculations. In general formula (13) is appropriate. But if the data are in whole number and their means are also in whole numbers, formula ;(14) is preferable. If the covariance and variance of the variable are known formula (16) is preferably used.

Substituting the values of a and b in the equation $\hat{Y} = a + b$

x we obtain the fitted regression as

$$Y = (\bar{y} - b\bar{x}) + bx$$

$$(\hat{y} - \bar{y}) = b(x - \bar{x}) \qquad \ldots\ldots\ldots (17)$$

The crown or hat (A) over y indicates that it is the estimated value of y, not the actual value. If we have obtained the numerical values of a and b, the same can be substituted in the equation Y = a + bx.

In the equation y = a +bx, a and b are known values. If we substitute the value of x for which Y is to be estimated, Y will be easily available.

The difference $(Y - \hat{Y})$ is known as the deviation of estimated value from actual value. These deviations are useful in testing of significance of regression parameters and correlation methods.

Regression line of *x* on Y

It has already been explained that similar to regression line of Y on *x,* there *is* a regression line x on, y whose equation can be given as.

$$X = \alpha_1 + \beta_1 Y + \varepsilon_1 \qquad \ldots\ldots\ldots (18)$$

Proceeding in the same manner as we did in case of regression line of Y on **X** the values of α_1 and β_1 can be estimated by the method of least squares. We leave the derivation, to the readers and directly give the values. An easy way to write all the formulas is that replace &Y by **X** and **X** by Y. Thus,

$$a_1 = (\bar{x} - b_1\bar{y}) \qquad \ldots\ldots\ldots (19)$$

$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum y_i^2 - \frac{(\sum y_i)^2}{n}} \quad \text{.....(20)}$$

for $i = 1,2,\ldots\ldots n$

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} \quad \text{.....(21)}$$

$$b = \frac{Cov(x,y)}{Var(Y)} \quad \text{.....(22)}$$

$$= \frac{\sigma x,y}{Var(Y)} \quad \text{.....(23)}$$

The regression equation of X on Y is,

$$(\hat{X} - \bar{x}) = b1(y - \bar{y}) \quad \text{.....(24)}$$

Here, for any given value of Y, X can be estimated.

Point of intersecticn of two regression lines

We have the regression line of Y on X as

$$(y - \bar{y}) = b(x - \bar{x})$$

and that of x on y as

$$(x - \bar{x}) = b1(y - \bar{y})$$

The point $(\bar{x}, \bar{y})$ satisfies both the above equation and hence the point of inte two lines is at the means of X and Y values. Since-two lines can intersect at one hence there can be no other point of intersection.

Regression coefficients in terms of correlation coefficient

Let us take the regression coefficient of Y on X β_{yx} as and that of x and y as β know that the correlations coefficient.

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The regression coefficients,

$$\beta yx = \frac{\sigma_{xy}}{\sigma_x^2}$$

and

$$\beta yx = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$\therefore \sigma xy = \rho \sigma_x \sigma_y$$

or
$$= \frac{\sigma_{xy}}{\sigma_x^2} = \rho \frac{\sigma_y}{\sigma_x}$$

$$\beta_{xy} = \rho \frac{\sigma_y}{\sigma_x} \qquad \text{.....(25)}$$

Again
$$\frac{\sigma_x y}{\sigma_y^2} = \rho \frac{\sigma_x}{\sigma_y}$$

or
$$\beta_{xy} = \rho \frac{\sigma_x}{\sigma_y} \qquad \text{.....(27)}$$

If we consider the estimated values,

$$b_{yx} = r \frac{S_y}{S_x} \qquad \text{.....(28)}$$

$$b_{xy} = r \frac{S_x}{S_y} \qquad \text{.....(29)}$$

Angle between the two lines of regression

If the equations of two straight lines are $y = m_1 x + c_1$ and $y = m_2 x + C_2$, the angle θ between the two lines is given as

$$\mathbf{Tan}\theta = \frac{\mathbf{m_1 - m_2}}{\mathbf{1 + m_1 m_2}}$$

Here our regression lines are,

$$Y = a + bxy\ X \qquad \text{.....(30)}$$

and $$X = a1 + bxy\ Y \qquad \text{.....(31)}$$

By equivalence,

and
$$m2 = bxy$$
$$m1 = 1/bxy$$

Hence, the tangent of the angle θ between the two regression lines is

$$\text{Tan}\,\theta = \frac{\frac{1}{bxy} - bxy}{1 + \frac{1}{bxy} bxy} = \frac{1 - byx\ \ bxy}{bxy + byx} \qquad \text{.....(32)}$$

Note: Here we have deliberately taken $m_1 = 1/b_{xy}$ and $m_2 = 1/b_{yx}$ as it makes the formula more convenient. Formula (32) in terms of r can be expressed in the following way

We know $$byx = r\frac{Sy}{Sx}$$

and $$bxy = r\frac{Sx}{Sy}$$

Substituting the values of b_{yx} and b_{xy} in (32) we get,

$$\text{Tan}\theta = \frac{1 - r\frac{Sy}{Sx} r\frac{Sx}{Sy}}{r - \frac{Sy}{Sx} + r\frac{Sx}{Sy}}$$

$$= \frac{(1-r^2)}{r}\left(\frac{Sx\ Sy}{S_y^2 + S_x^2}\right) \qquad \text{.....(33)}$$

$$\theta = \text{Tan}^{-1}\left[\frac{1-r^2}{r}\frac{(Sx\ \ Sy)}{S_x^2 + S_y^2}\right] \qquad \text{.....(34)}$$

$$r = 0, \text{Tan}\theta = \infty \text{ or } \theta = \frac{\pi}{2}$$

It shows that if the correlation between X and Y is zero, the two regression line are perpendicular to each other. If $r = \pm 1, Tan\theta = 0 or \theta = \pi$

So, when there is a perfect correlation between X and Y, the two lines of regression are either coincident or parallel to each other. Since both the lines pass through the point (X, Y) they cannot parallel. Hence, when there is a perfect correlation between the variables X and Y, the two lines of regression are coincident.

In terms of regression coefficients, if $b_{yx}\ b_{xy}, = 1 \text{Tan}\theta = \infty$ or $\theta = \pi$. Hence the two lines of regression are coincident.

If $b_{yx} = -b_{xy}$, Tan $\theta = \infty$ two regression lines are perpendicular to each other.

Properties of regression coefficient

Regression coefficients hold many properties which are divulged below.

(i) **Range**: The range of regression coefficient is $-\infty$ to ∞

(ii) **Signature property**: The sign of the regression coefficient b_{yx} and b_{xy} (β_{yx} and β_{xy}) and the correlation coefficient r(ρ) is the same. This is known as the signature property of regression coefficients. The reason is obvious. All the three coefficients have numerator Cov (X, Y) and denominator is positive. So if the Cov (x, y) is negative all the three coefficients are negative and if Cov (x, y) is positive, all of them are positive.

(iii) Fundamental property:

The geometric mean of the two regression coefficients is equal" to the correlation coefficient between the two variables. The sign of correlation coefficient will be the same as that of b_{yx} and b_{xy}. This is known as the fundamental property of regression coefficients. It is trivial to derive the relation. We know,

$$byx = r\frac{Sy}{Sx}$$

and

$$bxy = r\frac{Sx}{Sy}$$

Multiplying the two

$$byx \; bxy = r\frac{Sy}{Sx} \; r\frac{Sx}{Sy}$$

$$byx \; bxy = r^2$$

$$\text{or} \quad r = \pm\sqrt{byx \; bxy}$$

This mean b_{xy}, ρ and b_{yx} are in geometric progression.

(iv) Mean property:

The arithmetic mean of the positive regression coefficients is greater than the correlation coefficient between the two variables. This is known as the mean property of the regression coefficients notationally.

$$\frac{1}{2}(byx + bxy) > r$$

This relation can easily be proved. Consider the quantity

$$Q = \frac{1}{2}(byx + bxy) - \sqrt{byx\ bxy}$$

$$Q = \frac{1}{2}(byx + bxy) - \sqrt{byx\ bxy}$$

$$Q = \frac{1}{2}\left[byx + bxy - 2\sqrt{byx\ bxy}\right]$$

$$Q = \frac{1}{2}\left(\sqrt{byx} - \sqrt{bxy}\right)^2$$

Since $\left(\sqrt{byx} - \sqrt{bxy}\right)^2$ can never be negative, Q > 0. Q is zero if $\sqrt{byx} = \sqrt{bxy}$ but never negative.

(v) Magnitude property:

If one of the regression coefficients byx or bxy is greater than 1, the other is less than 1. This property of regression known as the magnitude property. The proof is very simple and direct. We know

$$byx\ \ bxy = r^2$$

$$r^2 < 1.$$

$$\therefore byx\ bxy < 1$$

$$byx < \frac{1}{bxy}$$

It is only possible when, $b^{xy} < 1$,then $b\mathbf{yx} > 1$ and vice - versa. They are equal only if byx = bxy =1

(vi) **Independence property**: If the two variables X and Y are independent, both the regression, efficient $= b_{Xy}$ = bxy =0 ,The reason is obvious that if x and y are independent Cov (x , *y)* = 0. Another way to prove that both the regression coefficients are zero is to express regression coefficients in terms of correlation coefficients. We know,

$$byx = r\frac{Sy}{Sx}$$

$$bxy = r\frac{Sx}{Sy}$$

If X and Y are independent, r = 0. If r = by x = 0 and bxy = 0.

Residual Variance

If Y is the estimated value of y from the regression equation $(y-\bar{y}) = b_{yx}(x - \bar{x})$, the residual variance is the expected value of the squares of the deviations of observed x values of y from the expected values of y as obtainable from the fitted line of regression i.e.

$$S_y^2 = E\ (y-\bar{y})^2$$

Substituting the value of, Y from the. regression equation We get,

$$S_y^2 = E\left[(y-\bar{y}) - b_{yx}(x-\bar{x})\right]^2$$

$$= E(y-\bar{y})^2 - b^2{}_{yx}E(x-\bar{x})^2 - 2b_{yx} - (y-\bar{y})(x-\bar{x})$$

$$= \sigma_y^2 + b_{yx}^2\sigma_x^2 - 2b_{yx} \quad \rho\sigma x \,.\, \sigma_y$$

$$= \sigma_y^2 + \rho^2\frac{\sigma_y^2}{\sigma_x^2}\sigma_x^2 - 2\rho\frac{\sigma_y}{\sigma_x}\rho\ \sigma x\ \sigma_y$$

$$= \sigma_y^2 - \rho^2\ \sigma_y^2$$

$$= \sigma_y^2(1-\rho^2)$$

If $= \rho = \pm,\ S_y^2 = 0$

If we consider the regression line of X on Y, we can derive in the same way that

$$S_y^2 = = \sigma_y^2(1-\rho^2)$$

Also $S_x^2 = 0$ if $\rho = \pm 1$.

This shows if $\rho = \pm 1$, the two lines of regression are coincident. If p departs from 1, the line of regression departs linearly. The quantity p (or r) is known as the coefficient of determination. Value of p or r near to one ensures linearity between the two variables and vice-versa. Also the quantity $(1 - \rho^2)$ *or* $(1 - r^2)$

is known as the coefficient of non-determination. Coefficient of correlation between observed and estimated value.

Consider the regression line of Y on X.

$$Y\ \bar{y} + b_{yx}(x - \bar{X})$$

$$= y + r\frac{\rho y}{\rho x}(x - \bar{x})$$

Let the observed value be denoted by Y. Some have to find out the correlation between Y and $\hat{Y}$ i.e. Cov (Y $\hat{Y}$).

By Pearson's Formula,

$$\text{Cov}\ (y,\hat{y}) = \frac{\text{Cov}(y,\hat{y})}{\sigma_y \qquad \sigma\hat{y}}$$

We know $= \hat{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

$$\therefore E(\bar{y}) \quad = \quad E(\bar{y}) + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$= \bar{y} + 0$$

$$= \bar{y}$$

$$\sigma 2^2_{\hat{y}} \quad = E[\hat{y} - E(y)]^2$$

$$= E[\hat{y} - \bar{y}]^2$$

$$= E\left[r\frac{\sigma y}{\sigma x}(x - \bar{x})\right]^2$$

$$= r^2\frac{\sigma_y^2}{\sigma_x^2}E(x - \bar{x})^2$$

$$= r^2 \frac{\sigma_y^2}{\sigma_x^2} E(x - \bar{x})^2$$

$$= r^2 \frac{\sigma_y^2}{\sigma_x^2} \sigma_x^2$$

$$= r^2 \sigma_y^2$$

$$= \therefore \sigma_y = r\ \sigma_y$$

Again

$$\begin{aligned}
\text{Cov}\ (y, \hat{y}) &= E[(y - E(y)(\hat{y} - E(\hat{y})] \\
&= E[byx(x - \bar{x})(y - \bar{y})] \\
&= byx\ E\left[(x - \bar{x}) r \frac{\sigma y}{\sigma x}(x - \bar{x})\right] \\
\therefore \text{Cou}(y, \hat{y}) &= byx\ \ r \frac{\sigma y}{\sigma x} E(x - \bar{x})^2 \\
&= byx\ \ r \frac{\sigma y}{\sigma x} \sigma_x^2 \\
&= r^2\ \frac{\sigma y^2}{\sigma x^2} \sigma \frac{2}{x} \\
&= r^2\ \frac{\sigma y^2}{\sigma x^2} \sigma \frac{2}{x}\ \text{since}\ byx = r \frac{\sigma y}{\sigma x} \\
&= r^2\ \sigma_y^2 \\
\therefore \text{Cou}(y, \hat{y}) &= \frac{r^2 \sigma_y^2}{r \sigma y \sigma x} \\
&= r
\end{aligned}$$

The result shows that the correlation coefficient between y and y is same as the correlation coefficient between x and y.

Example - 1.

Taking the data of example 1, chapter - 1 regarding test scores (x) and their sales performance we fit in the regression line of y on x and also estimates and sales performance of a sales representative who secure a score of 50 in the test. Also check whether the regression line is a good fit.

Also

(X) : 40 45 65 55 70 85 35 60 75 80

Scores

(Y) : 12 14 22 18 31 34 15 20 24 30

Sales (in lac Rs)

Firstly we find a: and

$$n = 10, \quad \Sigma x = 610, \quad \bar{x} = \frac{610}{10} = 61$$

$$\Sigma x = 220, \quad \bar{y} = \frac{220}{10} = 22$$

Since the means of x and y are whole numbers, we will use the following formula to calculate byx. byx $= \dfrac{\Sigma(xi - \bar{x})(yi - \bar{y})}{\Sigma(xi - \bar{x})^2}$

To calculate b_{yx}, we first prepare the following working table.

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$	$(y-\bar{y})$
40	12	-21	-10	441	210	100
45	14	-16	-8	256	128	64
65	22	4	0	16	00	00
55	18	-6	-4	36	24	16
70	31	9	9	81	81	81
85	34	24	12	256	288	144
35	15	-26	-7	746	182	49
60	20	-1	-2	1	2	4
75	24	14	2	196	28	4
80	30	19	8	361	152	64

$$byx = \frac{1095}{2390} = 0.458$$

Substituting the values $\bar{x}, \bar{y}$ of and b_{yx} in the regression equation

$$y = \bar{y} + byx(x - \bar{x})$$

We get the regression equation of y on x.

$$y = 22 + 0.458(x - 61) = 22 + 0.458x - 27.94 = -5.94 + 0.458x$$

When $X = 50$, the estimated value of y

$$y = -5.94 + 0.458 \times 50 = 16.94 \text{ lacs}$$

To check whether the regression line is a good fit, we calculate r^2. Firstly we calculate r by the formula.

$$r = \frac{\Sigma(xi - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(xi - \bar{x})^2 \Sigma(yi - \bar{y})^2}} = \frac{1095}{1121.22} = 0.977$$

$$\therefore r^2 = 0.95$$

Since the value of r^2 is very close to 1, we can say that the regression line is a good fit.

Example - **2**. We make use of the data of example - 2 of Lesson -13.

The data regarding years of service and ratings is as follows:

Employees	:	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Years of service	:	1	7	9	5	4	3	2	5	6	8	11	12	10	7	6	5	3	14
Ratings	:	5	8	6	3	7	5	4	6	9	10	7	8	9	3	5	2	9	5

The regression equation of y on x can be fitted in the following manner and also find whether the linear regression is a good fit or not.

First find $\bar{x}$ and $\bar{y}$

$$\Sigma xi = 118, \bar{x} = \frac{118}{18} = 6.56$$

$$\Sigma yi = 111, \bar{y} = \frac{111}{18} = 6.17$$

Since the mean value of y and x are not whole numbers, we would preferably use the following formula to calculate b_{yx}.

$$byx = \frac{\Sigma xiyi - (\Sigma xi)(\Sigma yi)/n}{\Sigma x_I^2 - \frac{(\Sigma xi)^2}{n}}$$

To work out various term in the formula for b_{yx}, we prepare the following table.

Employee	X	y	xy	x^2	y^2
A	1	5	5	1	25
B	7	8	56	49	64
C	9	6	54	81	36
D	5	3	175	25	9
E	4	7	28	16	49
F	3	5	15	9	25
G	2	4	8	4	16
H	5	6	30	25	36
I	6	9	54	3G	81
J	8	10	80	64	100
K	11	7	77	121	49
L	12	8	96	144	64
M	10	9	90	100	SI
N	7	3	21	49	9
0	6	5	30	36	25
P	5	2	If)	25	4
Q	3	9	27	9	81
R	14	5	70	196	25
Total	118	111	766	990	779

Making use of the partial calculations, the value

$$byx = \frac{776 - \frac{118 \times 111}{18}}{990 - \left(\frac{118^2}{18}\right)}$$

$$= \frac{766 - 727.67}{990 - 773.56}$$

$$= \frac{38.38}{216.44}$$

$$= 0.177$$

Thus, the regression equation of y on x is,

$$Y = \bar{y} + b_{yx}(x - \bar{x})$$

$$= 6.17 + 0.177(x - 6.56)$$

$$= 6.17 + 0.177x - 1.16$$

$$= 5.01 + 0.177x$$

To check whether the regression line is a good fit or not.

We calculate r by the formula.

$$r = \frac{\Sigma xiyi - \frac{(\Sigma xi)(\Sigma yi)}{n}}{\sqrt{\left[\Sigma x_1^2 - \frac{(\Sigma xi)^2}{n}\right]\left[\Sigma y_1^2 - \frac{(\Sigma yi)^2}{n}\right]}}$$

$$= \frac{766 - \frac{118 \times 111}{18}}{\sqrt{\left[990 - \left(\frac{118}{18}\right)^2\right]\left[779 - \left(\frac{111}{18}\right)^2\right]}}$$

$$= \frac{38.33}{\sqrt{216.44 \times 94.5}}$$

$$= \frac{38.33}{143.02}$$

$$= 0.27$$

$$\therefore r^2 = 0.0729$$

Since the value of r^2 is near to zero, we conclude that linear regression is not a good fit to the given data.

Example - 3.

For 12 pairs of observations, the following partial calculations are available.

$$\Sigma xy = 542, \Sigma x = 42, \Sigma y = 35, \sigma_x^2 = 20.6 \quad \text{and} \quad \sigma_y^2 = 16.4$$

Using the given calculation, we can find the two regression lines in the following manner.

First calculate Cov (x, y)

$$\text{cov}(x,y) = \frac{1}{n}\left[\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}\right]$$

$$= \frac{1}{12}\left[542 - \frac{42 \times 35}{12}\right]$$

$$\text{cov}(x,y) = \frac{1}{12}[542 - 122.5]$$

$$= \frac{419.5}{12}$$

$$= 34.96$$

$$\text{Thus,} \quad byx = \frac{\text{Cov}(x,y)}{\sigma_x^2}$$

$$= \frac{34.96}{20.6}$$

$$= 1.70$$

Similarly bxy $$= \frac{Cou(x,y)}{\sigma_y^2}$$

$$= \frac{34.96}{16.4}$$

$$= 2.13$$

Also, $x = \bar{x} = \frac{42}{12} = 3.5$ and $\bar{y} = \frac{35}{12} = 2.92$

The linear regression equation of yon x is

$$(\hat{y} - 2.92) = 1.70\ (x - 3.5)$$

$$\hat{y} = 1.70 \times -3.03$$

The regression line of x on y is

$$(\hat{x} - \bar{x}) = bxy\ \ (y - \bar{y})$$

$$\hat{x} - 3.5 = 2.13\ (Y - 2.92)$$

$$\mathbf{x = 2.13y - 2.72}$$

Example - 4 For n pair of values of x and y, the following results we found:

$r_{xy} = 6.5$, $\sigma_y = 8, \Sigma u^2 = 90$, $\Sigma uv = 120$ where $\mu = x - \bar{x}$ and $\upsilon = y - \bar{y}$. Find n, σ_x and the two regression coefficients.

$$\mathbf{b_{yx}} = \frac{\Sigma uv}{\sum u^2}$$

Solution:

$$= \frac{120}{90}$$

$$= \frac{4}{3}$$

We know,

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$\frac{4}{3} = 0.5\times\frac{8}{\sigma_x}$$

$$\therefore \sigma_x = 3$$

The formula for σx^2 is

$$\sigma x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$$

$$= \frac{1}{n}\Sigma u^2$$

$$9 = \frac{1}{n}\times 90$$

$$n = \frac{90}{9}$$

$$= 10$$

Again, $b_{xy} = \dfrac{\sum uv}{\sum v^2}$

$\sigma_y^2 = \frac{1}{n}\Sigma(y-\bar{y})$

$$= \frac{1}{n}\sum v^2$$

$$= 64 = \frac{1}{10}\sum v^2$$

$$\therefore \sum v^2 = 640$$

Thus $b_{xy} = \dfrac{120}{640}$

$$= \frac{3}{16}$$

Note: You can verify your answer. We know that

$$r = \sqrt{b_{yx} \quad b_{yx}}$$

$$r = \frac{4}{3} \times \frac{3}{16} = \frac{1}{2} = 0.5, \text{ which is the same value of r as given.}$$

Regression estimates through coding of data

By coding of data we mean-linear transformation of data. Under this process the origin of a variable is shifted and scale is changed. For instance if we are given the value Id, 20,30 40 and 50.Now we define a variable $\frac{x - 30}{10}$ Then for given values of x, the new variable dx has values

-2,-1,0,1,2 . In this process we note that x linearly relate with dx by the equation x-10 dx+ 30 . Also we have shifted the origin to 30 and on dividing by 10, the sale is reduced to one tenth. Such a transformation has reduced the data to a form which is easy to handle. Also Coding of data is very helpful, in fitting of trend in time series data In case of need we can retransform the value through the reverse process. Coding or linear transformation of data; not only saves time but also reduces the chances of error;

We discuss coding directly for the calculation of the intercept 'a' and regression coefficient 'b'

In regression analysis, we are concerned with two variables ac and y. Let a constant c_1 subtracted from x and C_2 from y. Then the reduced variables (x — c_1) and (y – c_2) are divided by d_1 and d_2 respectively. Thus, the coded values for x and y are

$$d_x \frac{x - c_1}{d_1}, dy = \frac{y - c_2}{d_2}$$

We have n pairs of values. $(x_1, y_1)\ (x_2, y_2) \ldots\ldots\ldots (x_n, y_n)$

For the ith pair (xi, yi); for $i = 1,2, \ldots\ldots n$. The coded values are.,

$$d_{xi} \frac{x_1 - c_1}{d_1}, dyi = \frac{yi - c_2}{d_2}$$

Again, the means of d_x and d_y can be given as follows:

$$d_x \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - c_1}{d_1} \right)$$

$$\frac{1}{n} \left[\frac{1}{d_1} \left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} c_1 \right) \right]$$

$$= \frac{1}{d_1} \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{d_1} . \frac{1}{n} n c_1$$

$$= \frac{1}{d_1} \bar{x} - \frac{1}{d_1} c_1$$

$$= \frac{\bar{x} - c_1}{d_1}$$

Similarly

$$dy = \frac{y - c_2}{d_2}$$

Now to calculate the regression. coefficient byx we prepare the following table.

x	y	$x-c_1$	$y-c_2$	$dx=\frac{x-c_1}{d_1}$	$dy=\frac{y-c_2}{d_2}$	dx dy	d_x^2	d_y^2
x_1	y_1	x_1-c_1	y_1-c_2	$dx_1=\frac{x_1-c_1}{d_1}$	$dy_1=\frac{y_1-c_2}{d_2}$	$dx_1 dy_1$	dx_1^2	dy_1^2
x_2 . . .	y_2	x_2-c_1	y_2-c_2	$dx_2=\frac{x_2-c_1}{d_1}$	$dy_2=\frac{y_2-c_2}{d_2}$	$dx_2 dy_2$	dx_2^2	dy_2^2
x_i	y_i	x_i-c_1	y_i-c_2	$dx_i=\frac{x_i-c_1}{d_1}$	$dy_i=\frac{y_i-c_2}{d_2}$	$dx_i dy_i$	dx_i^2	dy_i^2
x_n	y_n	x_n-c_1	y_n-c_2	$dx_n=\frac{x_n-c_1}{d_1}$	$dy_n=\frac{y_n-c_2}{d_2}$	$dx_n dy_n$	dx_n^2	dy_n^2
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n}(x_i-c_1)$	$\sum_{i=1}^{n}(y_i-c_2)$		$\sum_{i=1}^{n} dy_i$		$\sum_{i=1}^{n} d^2 x_i$	$\sum_{i=1}^{n} d^2 y_i$

Now we calculated the regression coefficient from the coded variables dx and dy and denote the regression coefficient of yon x from the coded variables as b_c. The formula for b_c is,

$$bc = \frac{\sum_i (dx_i - dx)(dy_i - dy)}{\sum_i (dx_i - dx)} \quad \ldots\ldots\ldots\ldots (38)$$

for i = 1,2,... ..,n

We know, $dx_i = \frac{xi - c_1}{d_1}$ and $dy_i = \frac{yi - c_2}{d_2}$

The means of the coded variables are,

$$\bar{dx} = \frac{\bar{x} - c_1}{d_1} \text{ and } \bar{dy} = \frac{\bar{y} - c_2}{d_2}$$

Substituting the values of dx_i and dy_i, $\bar{dx}$ and $\bar{dy}$ in (38) we obtain,

$$bc = \frac{\sum_{i=1}^{n}\left(\frac{xi-c_1}{d_1}-\frac{\bar{x}-c_1}{d_1}\right)\left(\frac{yi-c_2}{d_2}-\frac{\bar{y}-c12}{d_2}\right)}{\sum_{i=1}^{n}\left(\frac{xi-c_1}{d_1}-\frac{\bar{x}-c_1}{d_1}\right)}$$

$$= \frac{\sum_{i=1}^{n}\left(\frac{x_i}{d_1}-\frac{c_1}{d_1}-\frac{\bar{x}}{d_1}+\frac{c_1}{d_1}\right)\left(\frac{y_i}{d_2}-\frac{c_2}{d_2}-\frac{\bar{y}}{d_2}+\frac{c_2}{d_2}\right)}{\sum_{i=1}^{n}\left(\frac{x_i}{d_1}-\frac{c_1}{d_1}-\frac{\bar{x}}{d_1}+\frac{c_1}{d_1}\right)}$$

$$= \frac{\frac{1}{d_1 d_2}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{d_1^2}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{1}{d_2}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$bc = \frac{1}{d_2} byx \qquad \ldots\ldots\ldots\ldots (39)$$

The relation between b_c and byx in (39) reveals that the relation does not involve neither ci or C2. Hence, we can say that the change of origin does not affect the value of the regression coefficient. In a practical sense, we can say that we add or subtract a constant value from each X and also from Y, the two constants may or may not be different, the value of the regression coefficient remains the same.

The relation (39) between b_c and byx brings forth the fact that change of scale affect the value of the regression coefficient so, to get the original value of the regression coefficient from regression coefficient through coded data b_c, one should multiply by $\frac{d_2}{d_1}$

Following the same procedure, it is trivial to prove that

$$bc = \frac{1}{d_2} byx \qquad (40)$$

Note: 1.

The values of c_1, C_2, d 1 and dg are arbitrarily chosen looking to the data which result into the most convenient values of coded values. In a situation, they may be all different or, all the same, or some of them are the same the others differ.

2. The value of d_1 and d_2 should never be taken as zero, otherwise all variety values will

become infinite.

3. When $c_1 = 0$, $c_2 = 0$ it means no change of origin has been done.

4. When $d_1 = 1$, $d_2 = 1$, it means no change of scale has been implemented through coding.

5. If need be, coding can be confined to one variable only.

Example – 5

A company wanted to see the impact of advertising on sales of a re company collected the data which were as follows. (Fictitious data)

Advt. Expdt	(X)	:	6	12	18	24	30	36	42	48
(000' Rs.)										
Sales	(Y)	:	5	10	25	30	35	45	65	70
(Lac Rs.)										

(i). Find the regression line of Y on X.

(ii) Estimate the sales of the product of advertising worth Rs. 25,000.

Solution :

Fitting: of the regression .line can easily be done by using the method of coding. For this we prepare the following table by choosing $c_1 = 24$, $d_1 = 6$, $c_2 = 25$and $d_2 = 5$.

X	y	$dx = \frac{x - 27}{}$	$dy = \frac{y - 35}{5}$	dx dy	dx^2	dy^2
6	5	- 3	- 6	18	9	36
12	10	-2	-5	10	4	Z5
IS	25	- 1	-2	2	1	4
24	30	0	- 1	0	0	1
30	35	1	0	0	1	0
36	45	2	2	4	4	1
42	65	3	6	18	9	36
48	70	4	7	28	16	49
216	285	4	1	80	44	155

(i) The regression Coefficient from the coded values, can be calculated by the" formula,

$$bc = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{n}}{\sum d^2x - \left(\frac{\sum dx^2}{n}\right)}$$

$$= \frac{80 - \frac{4 \times 1}{8}}{44 - \frac{(4)^2}{8}}$$

$$= \frac{80 - 0.5}{44 - 2}$$

$$= \frac{79.5}{42}$$

$$= 1.89$$

The regression coefficient for the original data,

$$b_{yx} = \frac{d_2}{d_1} b_c$$

$$= \frac{5}{6} \times 1.89$$

$$= 1.575$$

Also $\bar{x} = \frac{216}{8} = 270$ and $\bar{y} = \frac{285}{8} = 35.625$

Hence regression line of y on x is

$$(y-35.625) = 1.575\ (x-27.0)$$

$$Y = -6.9 + 1.575\ x$$

(i) The estimated sales for $x = 25$

Example - 6. The following data relates to the number of scooters (in Lakhs) sold by & manufacturing company during the years 1985 to 1992.

Year	:	1985	1986	1987	1988	1989	1990	1991	1992
Number (in Lakhs)	:	6.0	6.1	5.2	5.0	4.6	4.8	4.1	6.2

Fit a straight line trend and estimate the sale for the year 199.3. (Take the year 1988 as working origin.)

Solution :

Since the years which are independent variate value cannot be used as such. Hence years are coded taking 1988 as origin. Such type of problems are usually faced with finding out the trend line. Let the trend line be Y = a + bx.

First prepare the computation table.

Year	X	Y	X^2	XY
1985	- 3	6.0	9	- 18.0
1986	-2	6.1	4	- 12.2
1987	- 1	5.2	1	-5.2
1988	0	5.0	0	00
1989	1	4.6	1	4.6
1990	2	4.8	4	9.6
1991	3	4.1	9	12.3
1992	4	6.2	16	24.8
Total	4	42.0	44	15.9

We can get the values of 'a' and 'b' in two ways. One is to substitute the value sof the terms in the formula for 'a' and V and the other is to write the normal equation and solve them directly for 'a' and 'b'.. Both lead to the same result.

By the formula,

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \left(\frac{\sum x}{n}\right)^2}$$

$$= \frac{15.9 - \frac{4 \times 42}{8}}{44 - \frac{(4)^2}{8}}$$

$$= \frac{15.9 - 21}{44 - 2}$$

$$= \frac{-5.1}{42}$$

$$= -0.121$$

$$\bar{y} = \frac{42}{8} = 5.25, \quad \bar{x} = \frac{4}{8} = 0.5$$

$$a = \bar{y} - b\bar{x}$$

∴ The required trend line is

$$y = 5.31 - 0.121x$$

Secondly if we solve the normal equations

$$y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Substituting the values from the table, the normal equation come out to be,

$$42 = 8a + 4b$$

$$15.9 = 4a + 44$$

The readers can solve these equation and verify that the same values of 'a' and 'b are obtained. Curvilinear Regression : The relation between the dependent and independent variable (s) is not necessarily linear. It is often curvilinear. The curve may be of any type, a parabola, a cubic, a polynomial of degree K, an as trod, a hyperbola or any other type of curve. But confining to the requirement of the course, we will consider some well known curves which can be reduced to linear relations under transformation.

Exponential growth curve

Mathematical equation of the curve is,

$$Y = \alpha\beta^x \qquad \ldots\ldots\ldots\ldots\ldots\ldots(41)$$

If we put = 1 + *i* where 'i' is the rate of interest and x, the number of year than the equation of the growth curve reduces to the formula for compound interest and a if the initial amount invested. Obviously, y is the amount which will swell under compound interest in x years. It is not only the question of business only but many scientific data also follow this law. Population growth is one of them.

In equation (41), if we put x = 0, then y = *a* (0 , *a)* is the point on the 'y' from where the curve starts.

The filling of exponential growth curve as such appears to be a tedious problem. But it becomes trivial as soon as we take logarithm of both the side. The equation' becomes linear in log terms I.e.

$$\text{Log } Y = \log a + X \text{ Log } \beta \qquad \ldots\ldots\ldots\ldots(42)$$

Here we note that the log term appears only for Y but not X. If we put log Y = z, log a — a and log = b , the equation takes the form,

$$Z = a + bx \qquad \ldots\ldots\ldots\ldots(43)$$

So in the fitting of the growth curve, first the variate values of Y are to be transformed to log - values and then linear equation z - a + bx is fitted in the same manner as we do for the regression equation of Y on X. Here it is a simple linear regression of z on X. The formulae for a and b in the like manner for n paired values (x_1, z_1) , $(x_2 , z_)$,...(xn, z_n) can be given as

$$a = (\bar{z} - b\bar{x}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots (44)$$

And

$$b = \frac{\sum_{i=1}^{n} xi\ zi - \frac{\left(\sum_{i=1}^{n} xi\right)\left(\sum_{i=1}^{n} zi\right)}{n}}{\sum_{i=1}^{n} x_1^2 - \frac{\left(\sum_{i=1}^{n} x_1^2\right)}{n}}$$

for i = 1,2,... ... .n

Once we get the values of a and b, we can get the values of a and β by taking antilog of a and b respectively. Substituting the estimated values of α and β in the equation

$Y = \alpha\beta^x$ we get the fitted regression equation, $y = ab^x$

Example 7 The data given below show the atmospheric pressure at various heights above the sea level.

S.No.	Heights (Kms)	Pressure (Cms.)	S.No	Heights (Kms)	Pressure (Cms.)
1.	0	73.4	11.	4.7	21.6
2.	1.3	68.3	12.	8.7	25.4
3.	2.3	51.8	13.	9.0	19.3
4.	2.9	55.6	14.	10.9	17.5
5.	4.2	48.5	15.	11.4	15.7
6.	4.2	41.1	16.	11.8	12.2
7.	5.3	33.5	17.	12.9	14.2
8.	6.0	34.8	18.	13.5	10.4
9.	6.4	28.4	19.	13.5	11.7
10.	7.6	28.4	20	15.3	9.4

Fit in the exponential curve $y = \alpha\beta^x$

Solution : To fit in the exponential curve we prepare the following computation table:

S.No.	Height (in Kms)	Atmos. Pressure (in Cms.)	LogY = Z	ZX	X^2
1	0	73.4	1.866	00	00
2	1.3	68.4	1.834	2.3842	1.69
3	2.3	51.8	1.714	3.9422	5.29
4	2.9	55.6	1.745	5.0605	8.41
5	4.2	48.5	1.686	7.0812	17.64
6	4.2	41.1	1.614	6.7788	17.64
7	5.3	33.5	1.525	8.0825	28.09
8	6.0	34.8	1.542	9.2520	36.00
9	6.4	28.4	1.453	9.2992	40.96
10	7.6	28.4	1.453	11.0428	57.76
u	4.7	21.6	1.334	6.2698	22.09
12	8.7	25.4	1.405	12.2235	75.69
13	9.0	19.3	1.286	11.5740	81.00
14	10:9	17.5	1.243	13.5487	118.81
15	11.4	15.7	1.196	13.6340	129.96
16	11.8	12.2	1 .086	12.8148	139.24
17	12.9	14.2	1.152	14.8608	166.41
IS	13.5	10.4	1.017	13.7295	182.25
19	13.5	11.7	1.068	14.4180	182.25
20	15.3	9.4	0.973	14.8869	234.09
Total	151.9		28.192	190.8834	1545.27

$$\bar{z} = \frac{28.192}{20} = 1.4096, \quad \bar{x} = \frac{151.9}{20} = 7.595$$

By the formula

$$b = \frac{190.8834 - \frac{(151.9)(28.192)}{20}}{1545.27 - \frac{(151.9)^2}{20}}$$

$$= \frac{190.8834 - 21411.84}{1545.27 - 1153.68}$$

$$= \frac{-23.235}{391.59}$$

$$= -.0593$$

$$\bar{\beta} = 0.872$$

$$a = 1.40096 - (-.0593)(7.595)$$

= 1.4096 + 0.4504

= 1.86

Taking antilog,

$$\hat{\alpha} = 72.44$$

Hence the equation of the exponential curve is,

$$Y = (72.44)\ (0.872)$$

As a check of our equation, we estimate Y for some value of X given in the table.

$X = 0, \quad \hat{Y} = 72.44 \quad Y = 73.4$

$X = 9, \quad Y = 19.3$

$\hat{Y} = 21.12$

$X = 6 \quad Y = 34.8$

$\hat{Y} = 31.85$

The observed and estimated values show that the curve is a good fit.

Logarithmic curve

The mathematical equation of the logarithmic curve is,

$$Y = \alpha X^{\beta} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(46)$$

If we take logarithm of the equation, (46), it becomes linear equation in log x and log y. All the more, both the, variable occur in log terms, that is why, the curve $Y = \alpha X^{\beta}$ is known as logarithimic curve. Thus

$$\log Y = \log \alpha + \beta \log X \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(47)$$

From equation (46), it is evident

When $X = 0$, $Y = 0$, but equation (47) makes us not to choose $X = 0$ because $\log 0 = -\infty$ and it is not possible to trace the curve. Hence, it is better to take initial value of X say, $X = 1$, The curve fakes positive slope' when β is positive. Again when β is negative, the curve is negatively sloping shown by the broken lines. If we put $\log Y = Z$ and $\log X = u$, $\log a = y$ Then the equation reduces to the linear equation in z and u.

$$Z = \gamma + \beta u \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(48)$$

Let the estimated values of y and b be a and b respectively. The formulae for a and b for n paired observations can be given as

$$a = \bar{z} - b\bar{u} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(49)$$

$$\text{and } b = \frac{\sum_{i=1}^{n} zi\ ui - \frac{\left(\sum_{i=1}^{ni} z\right)\left(\sum_{i=1}^{n} u\ i\right)}{n}}{\sum_{i=1}^{n} u_1^2 - \frac{\left(\sum_{i=1}^{n} u\ i\right)^2}{n}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(50)$$

Taking antilog, of 'a' we obtain the estimated values of a and using the value of 'b' as such Substituting the values of a and b, the logarithmic curve is,

$$Y = a X^{h} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(51)$$

Example • 8 The following observations were recorded in an astronomical study with regard to the distance of planets from the sun and periods of revolution.

Planets	Distance Astronomical(X)	Periods in Years(Y)
Mercury	0.39	0.24
Venus	0.72	0.62
Earth	1.00	1.00
Mars	1.52	1.88
Jupiter	5.20	11.9
Saturn	9.54	29.5
Uranus	19.20	84.0
Neptune	30.10	165.0
Pluto	39.40	248.0

Fit the logarithmic curve.

Solution : To fit in the logarithmic curve $Y = \alpha X^{\beta}$ we first prepare the following computation table.

Planets	Distance (X)	Periods (Y)	Log X = u	Log Y = 2	uZ	u^2
Mercury	0.39	0.24	- 0.4089	- 0.6198	0.2534	0.1672
Venus	0.72	0.62	- 0.1427	- 0.2076	0.2096	0.0204
Earth	1.00	1.00	00	00	00	00
Mars	1.52	1.88	0.1818	0.2746	0.0499	0.0330
Jupiter	5.20	11.9	0.7160	1.0755	0.7700	0.5126
Saturn	9.54	29.5	0.9795	1.4698	1.4397	0.9594
Uranus	19.20	84.0	1.2833	1.9242	2.4693	1.6468
Neptune	30.10	165.0	1.4786	2.2175	3.2788	2.1862
Pluto	39.40	248.0	1.5955	2.3944	3.8202	2.5456
Total	107.07	542.14	5.6831	8.5286	12.1109	8.0712

Now,

$$\bar{x} = \frac{107.07}{9} = 11.99, \quad \bar{U} = 0.6314,$$

$$\bar{x} = \frac{542.14}{9} = 60.24, \quad \bar{U} = 0.9476$$

By the formula (50) and (51)

$$b = \frac{12.1109 - \frac{(8.5286)\times(5.6831)}{9}}{8.0712 - \frac{(5.6831)^2}{3.5886}}$$

$$= \frac{12.1109 - 5.3854}{8.0712 - 3.5886}$$

$$= \frac{6.7255}{4.4826}$$

$$= 1.50$$

$$\text{and } a = 0.9476 - 1.5 \times 0.6314$$

$$= 0.9476 - 0.9471$$

$$= 0.0005$$

$$a = \log \hat{a} = 0.0005$$

Taking antilog,

$$\hat{a} = 1.0011$$

Thus, the estimated logarithmic curve is,

$$Y = (1.0011)(X)^{1.5}$$

Just to verify the exactness of the fitted curve, estimate Y for a few values of X. when,

x = 1,	Y = 1.0011
x = 5.2,	Y = 11.87
x = 0.72	Y = 0.61
x = 30.10	Y = 165.32

The comparison of estimated values with the observed values reveals that the fitted logarithmic curve is a very good, fit to the data.

The Reciprocal curve

The mathematical equation of the curve is

$$\frac{1}{Y} = \alpha + \beta x \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (52)$$

When $X = 0 \frac{1}{Y} = a$ Also when the value of b in the equation (52) is positive, the curve has a negative slope and vice versa.

If we substitute $\frac{1}{Y} = z$ the equation (52) becomes linear in Z and X.

$$Z = a + \beta x \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (53)$$

Using the formulae for estimates of α and β as a and b respectively, the formulae

$$a = \bar{z} - b\bar{x}$$

are

and $$b = \frac{\sum zi\ xi - \frac{\left(\sum_i xi\right)\left(\sum_i zi\right)}{n}}{\sum x1^2 - \left(\sum x1\right)^{2/n}}$$

Other details remain the same as with logarithmic or semilogarithmic curves.

Example - 9

The production of potato (Million Tonnes) and prices per quintal during the last thirteen years are as follows :

Year	Production of Potato (Million Tonnes)	Price(per Kg.)
1981	2.7	1.6
1982	2.6	2.0
1983	2.3	2.6
1984	3.2	12
1985	2.7	2.0
1986	2.5	3.1
1987	3.5	1.3
1988	2.7	2.1
1989	3.5	1.4
1990	2.8	2.0
1991	2.2	3.9
1992	3.2	1.9
1993	3.0	1.8

$$\frac{1}{Y} = \alpha + \beta x$$

We first fit in the transformed equation as a straight line by taking $\frac{1}{Y} = z$ For the purpose we first prepare the following computation table.

Year	Production of Potato (M. Tonnes) (X)	Prices Per Kg	$\frac{1}{y} = z$	$X\bar{z}$	x^2
1981	2.7	1.6	0.625	1.6875	7.29
1982	2.6	2.0	0.500	1.3000	6.76
1983	2.3	2.6	0.385	0.8855	5.29
1984	3.2	1.2	0.833	2.6656	10.24
1985	2.7	2.0	0.500	1.3500	7.29
1986	2.5	3.1	0.322	0.8050	6.25
1987	3.5	1.3	0.769	2.6915	12.25
1988	2.7	2.1	0.476	1.2852	7.29
1989	3.5	1.4	0.714	2.4990	12.25
1990	2.8	2.0	0.500	1.4000	7.84
1991	2.2	3.9	0.256	0.5632	4.84
1992	3.2	1.9	0.526	1.6832	10.24

By the formula (55)

$$b = \frac{20.4837 - \frac{(36.9)\times(6.962)}{13}}{106.83 - \frac{(36.9)^2}{13}}$$
$$= \frac{20.4837 - 19.7614}{106.83 - 104.74}$$
$$= \frac{0.7223}{2.09}$$
$$= 0.3456$$

and by the formula (54),

$$a = 0.5355 - 0.3456 \times 2.838$$
$$= 0.5355 - 0.9808$$
$$= -0.4453$$

Where, $\bar{x} = \frac{36.9}{13} = 2.838$

and $\bar{z} = \frac{60962}{13} = 0.5355$

Thus, the equation of the reciprocal curve is,

$$Z = -0.4453 + 0.346\chi$$

Putting $Z = \frac{1}{Y}$

The requited reciprocal curve is,

$$\frac{1}{Y} = 0.4453 + 0.346\chi$$

Alternative method

If the reader want, they can calculate the value of a and b directly by,solving **the** normal equation:

$$13a + 36.9b = 6.962$$
$$39.9a + 106.83b = 20.4837$$

It is trivial to verify that,

$$a = -0.4453$$
$$b = 0.5355$$

QUESTIONS

1. In what respects, a statistical model differs from mathematical model.
2. What are the assumptions made in a linear regression model ?
3. Why do we get two regression equations ?
4. When are the two regression equations identical ?
5. Define regression coefficient of Y on X.
6. Give least squares method of estimation of regression parameter in a simple linear regression equation.
7. Write the properties of regression coefficient(s).
8. In what way the regression coefficients are connected with correlation coefficients.
9. How the two regression coefficients are connected with each other in respect of magnitudes.
10. Find the angle between two regression lines.
11. Define and discuss coefficient of determinate.
12. Derive the correlation between Y and Y, the least square regression estimate of Y.

13. A building contractor is interested in knowing whether relationship does not exist between the number of building permits issued and the volume of sales of such buildings in some past years. He collects data about sales (Y, in Thousands of rupees) and the number of building permits issued (X, in hundreds) in past 10 years. The results worked out are as under :

$$\sum x = 117, \sum y = 78, \sum xy = 981 \sum x^2 = 1491, \sum y^2 = 662$$

(i) What level of sales can you expect next year, it it is hoped that 20Q0 building permits would be issued?

(ii) What change in sales is likely to take place with an increase of 100 building permits?

14. Fit the curve $Y = ax^b$ to the following data

x	:	1	2	3	4	5	6
y	:	1200	900	600	200	110	50

15 Fit a straight line trend to the following data :

Year	:	1984	1985	1986	1987	1988	1989	1990
Production ('000 Tonnes)	:	37	38	37	40	41	45	50

Estimate the production for the year 1992.

16. An investigator reported that $b_{yx} = 4.3$ and $b_{yx} = 0.6$ comment on the values of b_{yx} and b_{yx}

17. The following table gives aptitude test scores and productivity indices of 8 randomly selected workets:

Aptitude Scores	:	57	58	59	59	60	61	62	64
Productivity Index	:	67	68	65	68	72	72	69	71

Estimate the productivity Index of a worker whose test score is 65.

18. Given the two regression lines,

3X +2y = 12

5x + y = 13

(i) Find which line out of the two represent the regression line of y on x to of x on Y.

(ii) Find the correlation coefficient between X and Y.(Hi) At what point the two lines intersect.

(iii) At what point the two lines intersect.

(iv) What is the regression estimate of y for x = 0.

19. The area under tea cultivation from 1945 to 1951 is given below

Year	Area ('000 hectares)
1945	670
1946	680
1947	690
1948	700
1949	710

1950	720
1951	730

(i) Fit in the trend line

(ii) Estimate the area under tea cultivations in 1955.

- End Of Chapter -

LESSON -15

TESTING AND INTERVAL ESTIMATION OF REGRESSION COEFFICIENT

Preamble

The fitting of a regression-equation is done to estimate the dependent variable Y through the independent variable X. The regression parameters a, and b play the complete role in estimation process. When regression analysis is done, some values of the parameters a, and b are obtained. Now it becomes essential to know whether the contribution of these parametric, values in estimating Y is significant or not. If the value of b is non-significant, then it shows that the estimation of Y through X is not meaningful. Since if b = 0, r = 0, it means that the two variable are not linearly retard. -Hence Simple linear equation is not fit for estimating Y through X. For instance, if we want to estimate sale of a product on the basis of test scores of the salesman. If IT comes out to be non-significant, there is no sense in relating the sales with the test spores of the salesman. Also if a 0, it means that the line passes through the origin, In this way, there is no intercept in reality and its contribution in estimating Y is insignificant. Therefore, one should test the significance of regression parameters prior to estimating the variate values.

Test of significance of regression coefficient

Test of significance of b_{yx}, the regression coefficient of Y on X amounts to testing.

$H_o : \beta yx = o$ us $H1 : \beta yx \neq 0$

Test of significance of β_{yx}, the regression coefficient of Y on X amounts to testing.

Let the estimation of β_{yx} is based on n paired values $(x_1\ y_1)$, (X_2,y_2),... (x_n, y_n) The hypothesis H_0 against H_o can be tested by t-test. The test statistic is,

$$t = \frac{b}{sb} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

t has(n-2)d.f

Whereas, b is the least square estimate of β_{yx} and sb is the standard deviation of b. As we know the value of b is obtained by the formula,

$$b = \frac{\sum_{i=1}^{n} xi\ yi - \frac{\sum_{i=1}^{n} xi \sum_{i=1}^{n} yi}{n}}{\sum_{i=1}^{n} x_1^2 - \frac{\left(\sum_{i=1}^{n} xi\right)^2}{n}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

for i =1,2,.....n.

To calculate Sb, we first calculate Se where,

$$S_e^2 = \frac{1}{(n-2)} \sum_{i}^{n} (yi - a - bxi)^2 \qquad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

The divisor (n-2) in Se^2 is used because two parameters α and β are estimated resulting into the loss of two d.f. All the more Se^2 obtained by the formula (3) is an unbiased estimate of σ_e^2 S_e^2 also known as *mean squared error*.

With a little algebraic manipulation, it. is easy to show that

$$S_e^2 = \frac{1}{(n-2)}\left\{\sum_i^n (yi-\bar{y})^2 - b\sum_i^n (xi-\bar{x})(yi-\bar{y})\right\} \qquad \ldots\ldots\ldots\ldots\ldots(4)$$

Suppose $ui = xi$ and $vi = (yi-\bar{y})$

So the formula for Se in (4) changes **to**,

$$S_e^2 = \frac{1}{n-2}\left\{\sum_i v_1^2 - b,. \sum_i uivi\right\}$$

Now putting the value of b as $\sum_{i=1}^n uivi / \sum_{i=1}^n u_1^2$

$$S_e^2 = \frac{1}{(n-2)}\left[\sum_{i=1}^n u_1^2 - \frac{\left(\sum_{i=1}^n uivi\right)^2}{\sum_{i=1}^n u_1^2}\right] \qquad \ldots\ldots\ldots\ldots\ldots(4.2)$$

Once we know Se^2 , it is easy to find Sb^2 and thereby Sb Thus,

$$S_b^2 = \frac{S_e^2}{\sum_{i=1}^n (xi-\bar{x})^2}$$

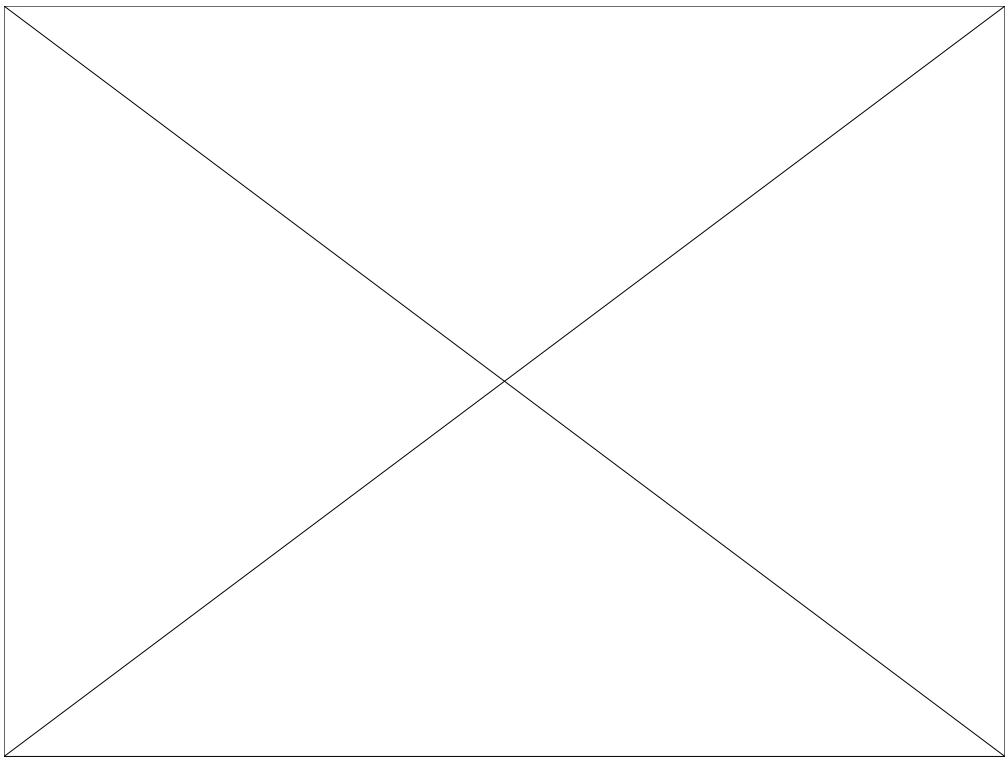
$$= \frac{S_e^2}{\sum_{i=1}^n u_1^2}$$

Again,

$$Sb = \sqrt{S_b^2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(6)$$

Substituting the value of b and Sb in (I) we get the calculated value of t.

To make a decision about Ho, compare the calculated value of t With the tabulated value of t for (n-2) degrees of freedom and a level of significance. The decision criteria is, reject Ho if t cal > t a , (n-2) and otherwise accept Ho. Accepting Ho means that the' change in Y corresponding to an unit change in X is of no significance.



TEST OF SIGNIFICANCE OF THE INTERCEPT

Test of significance of the intercept amounts to testing

$$H_o : \alpha = 0 \text{ Vs } H_1 : \alpha \neq 0$$

The test statistics for testing H_0 is,

$$t = \frac{a}{S_a} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

t has (n-2) d.f

Where, $a = (\bar{y} - b\bar{x})$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

and Se, the standard deviation of 'a' can be obtained by the formula,

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (9)$$

$$S_a^2 = S_e^2\left\{\frac{1}{n} + \frac{\overline{X}^2}{\Sigma i(xi - \overline{x})^2}\right\}$$

$$S_a^2 = S_e^2\left\{\frac{1}{n} + \frac{\overline{X}^2}{\Sigma i\, u_1^2}\right\} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9.1)$$

We have already known $Se^2\overline{X}^2$ and Σui^2. Hance it is very easy to calculate Sa^2.

Also, $Sa = +\sqrt{S_a^2}$

The decision about Ho is taken in the same manner as we do in case of testing $\beta_{yx} = 0$, i.e reject Ho if t cal >t α (n-2) and otherwise accept Ho.

Note: The test procedure in case of regression line of x on Y, remains the same > except that replace X by Y and Y by X or u by v and v by u.

Analysis of variance technique for testing the significance of Regression coefficient

The hypothesis,

$$H_o : \beta_{yx} = o \text{ vs. } H_1 : \beta_{yx} \neq 0$$

can be tested by F-test also with the help of analysis of variance (ANOVA) table.

ANOVA Table

Source	d.f.	Sum of Squares	Mean Sum of Squares	F-value
Regression	1	$b \ \Sigma_i u_i v_i$	$\left(b\sum_{i=1}^{n} ui \ vi \,/1\right)$	$\left(b\sum_{i=1}^{n} ui \ ui\right)/S_a^2 = F$
Deviation from regression	(n-1)	$\sum_{i=1}^{n} u_1^2 - b\sum_{i=1}^{n} ui \ ui$	$\frac{\sum_{i=1}^{n} u_1^2 - b\sum_{i=1}^{n} ui \ ui}{n-1} = S_e^2$	
Total	(n-2)	$\sum_{i=1}^{n} u_1^2$		

To decide about Ho,, compare the calculated value of F with the tabulated value of for (1, n-2) d.f. and a level of significance. If F cal > F a , (n-2), reject Ho. It means that the regression coefficient H_0 is significant and thus makes a significant contribution in estimating Y through X

If F cal<F a, n-2, Ho is accepted.

Note: Level of significance has to be decided prior to investigations. But in practice it is mostly taken as a = 6.05 or a = 0.01 if nothing is prefixed. If a is chosen as 0.01, it is called highly significant level.

Example-1: The following data provides the information about the capital invested in crore rupees and the profit earned in crore rupees.

Capital	:	100	90	85	75	60	45	40	35	20
Profit	:	25	20	18	15	13	8	4	3	2

(i) Fit in the regression line of profit on capital.

(ii) Estimate Y for X = 50

(iii) Test the significance of the recession coefficient, (iv) Test the significance of the intercept.

To fit in the regression line of profit (y) on capital (X), we prepare the following computation table.

Capital X	Profit Y	$3C^2_i$	Y^2_i	xy
100	25	10000	625	2500
90	20	8100	400	1800
85	18	7225	324	1530
75	15	5625	225	1125
60	13	3600	i69	750
45	8	2025	64	360
40	4	1600	16	160
35	33	1225	8	105
20	2	400	4	40
550	108	39800	1836	8370

Other values required for fitting the regression equation and testing are computed in the following manner.

$$\Sigma x = 550, \quad \overline{X} = \frac{550}{9} = 61.11$$

$$\Sigma y = 108, \quad \overline{y} = \frac{108}{9} = 12.00$$

$$\Sigma ui\ vi = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{N}$$

$$= 8370 - \frac{550 \times 108}{9}$$

$$= 8370 - 6600$$

$$= 1770$$

$$\Sigma u_1^2 = 39800 - \frac{(550)^2}{9}$$

$$= 1836 - 1269$$

$$= 540$$

$$\text{Thus}, b = \frac{\Sigma ui\ vi}{\Sigma u_1^2}$$

$$= \frac{1770}{6188.89}$$

$$= 0.286$$

$$a = \overline{Y} - b\overline{x}$$

$$= 12.00 - 0.286 \times 61.11$$

$$= -5.48$$

Thus, the regression line is

$$Y = -5.48 + 0.286X$$

(ii) When $X = 50$

$$= 8.82$$

(iii) To test

$$H_o : \beta_{yx} = o \text{ vs } H1 : \beta_{yx} \neq 0$$

$$S_e^2 = \frac{1}{n-2}\{\Sigma u_1^2 - b\Sigma i\ ui\ vi\}$$

$$= \frac{1}{9-2}\{540 - 0.286\ x\ 1770\}$$

$$= \frac{1}{7}\{540 - 506.22\}$$

$$= \frac{33.78}{7}$$

$$= 4.82$$

$$S_t^2 = S_e^2 / \Sigma u_1^2$$

$$= \frac{4.82}{6188.89}$$

$$= 0.00078$$

$$\therefore Sb = 0.0279$$

The Statistic,

$$t = \frac{0.286}{0.0279}$$

$$= 10.25$$

t has 7 d.f.

Tabulated value oft for 7 d.f. and 5 % level of significance for two tailed test is 2.365.Since t cal >2.365, we reject Ho. It means that the regression coefficient is significant.

To test the hypothesis,

$$H_o : a = o \text{ vs } H1 : a \neq 0$$

We calculate S_a^2=The test statistics,

$$\begin{aligned} Sa^2 &= S_e^2\left\{\frac{1}{9}+\frac{(61.11)^2}{6188.89}\right\} \\ &= 4.8\left\{\frac{1}{9}+\frac{3734.43}{6188.89}\right\} \\ &= 4.8\{0.1111+0.6034\} \\ &= 4.8\times 0.7145 \\ &= 3.44 \\ \therefore Sq &= 1.85 \end{aligned}$$

The test statistic

$$\begin{aligned} t &= \frac{-5.48}{1.85} \\ &= -2.96 \end{aligned}$$

{t cal } > 2.365 (the tabulated value of t for 7 d.f. and $\alpha = 0.05$), we reject H_0. It means that the intercept has a significant value vis-a-vis it plays a significant role in estimating Y through X.

Example-2: Given the following paired set of 7 observation on the production of a commodity from 1985 to 1991.

Year	:	1985	1986	1987	1988	1989	1990	1991
Production	:	8	10	12	11	14	18	22

(i) Taking 1988 as origin, fit in the trend line y = a + bx

(ii) Estimate the production for the year 1993.

(iii) Test the significance of the regression coefficient.

Solution:

To fit in the trend line and test of hypothesis we prepare the following computation table.

Year	coded Years X	Production Y		XY	Y^2
1985	-3	8	9	-24	64
1986	-2	10	4	-20	100
1987	-1	12	1	-12	144
1988	0	11	0	0	121
1989	1	14	1	14	196
1990	2	18	4	36	324
1991	3	22	9	66	484
Total	0	95	28	60	1433

With the help of the above computations,

$$\overline{X} = 0, \quad \overline{Y} = \frac{95}{7} = 13.57$$

$$\Sigma u_i^2 = 28, \Sigma u_i^2 = 1433 - \frac{(95)^2}{7}$$

$$= 1433 - 1289.28$$

$$= 143.72$$

$$\Sigma iuivi = 60 - \frac{0 \times 95}{7} = 60$$

To fit in the trend line we, compute

$$b = \frac{\Sigma uivi}{\Sigma u_i^2}$$

$$= \frac{60}{28}$$

$$= 2.14$$

$$a = \overline{Y} - b\overline{x}$$

$$= 13.57 - 2.14 \times 0$$

$$= 13.57$$

The trend line is

Y= 13.57 + 2.14 X

i. To estimate the production for 1993

X = 5, hence

$$\begin{aligned} Y &= 13.57 + 2.14 \times 5 \\ &= 13.57 + 10.7 \\ &= 24.27 \end{aligned}$$

ii. to test the significance of the regression coefficient We calculate,

$$\begin{aligned} S_e^2 &= \frac{1}{n-2}\left(\Sigma u_1^2 - b\,\Sigma uivi\right) \\ &= \frac{1}{7-2}\left(143.72 - 2.14 \times 60\right) \\ &= \frac{1}{5}\left(143.72 - 128.4\right) \\ &= \frac{15.32}{5} \\ S_b^2 &= S_e^2 / \Sigma u_1^2 \\ &= \frac{3.064}{28} \\ &= 0.1094 \\ &= 0.33 \\ \therefore Sb &= 0.574 \end{aligned}$$

The test statistics,

$$\begin{aligned} t &= \frac{b}{Sb} \\ &= \frac{2.14}{0.573} \\ &= 3.73 \end{aligned}$$

Tabulated value of t for 5 d.f. and $\alpha = 0.05$ is 2.571. Hence t cal > 2.571. We reject Ho which means that § is significant. This reveals that the variable X makes a significant contribution in estimating Y.

Interval Estimation

The confidence limits for a parameter 6 corresponding to the confidence probability (1 - a) are given by the formula,

The confidence limits for a parameter δ corresponding to the confidence probability $(1 - \alpha)$ are given by the formula,

$$\theta \mu x^{1} S\hat{\theta} \qquad \ldots\ldots\ldots\ldots\ldots (11)$$

In formula (11) $\hat{\theta}$ is the estimated value of θ, x' is the deviate value for the distribution f (x, $\hat{\theta}$) and S0 is the standard error of $\hat{\theta}$. The lower limit for θ is $\hat{\theta}$ - X S $\hat{\theta}$ and upper limit is $\hat{\theta}$ +x' $S\hat{\theta}$ By analogy, the confidence limits for βyx are given by the formula,

$$b \mu Sb \; t \; a \; (n-2)$$

If formula (12), b is the estimated value of β and S_b is the standard error of b and t a (n-2) is the tabulated value of t for *ft* level of significance and (n-2) d.f. substituting the values of b, Sb and t a , (n-2) we obtain the lower limit by the formula b- Sb ta, n-2 and upper limit by b + Sb. *ta* , n-2.

Example-3 : Given the following results:

$$n = 9, \bar{x} = 5, \bar{y} = 12, \sigma x = 3.0, r = 0.80$$

(i) Fits in the two regression lines of y on x.

(ii) Test the significance of the regression coefficients.

(iii) Test the significance of the incepts of the two lines.

(iv) Find 95% confidence interval for the regression coefficient

Solution :

We first make the following computations.

$$\sum uivi = r\sigma x\sigma y$$

$$= .8 \times 2.6 \times 3.0$$

$$= 6.24$$

Similarly,

$$\sigma_y^2 = \frac{1}{9}\Sigma u_1^2$$

$$9 = \frac{1}{9}\Sigma u_1^2$$

$$\Sigma u_1^2 = 81$$

$$\sigma_x^2 = \frac{1}{9}\Sigma u_1^2 \text{ or } 6.76 = \frac{1}{9}\Sigma u_1^2 \text{ or } \Sigma u_1^2 = 60.84$$

Again, suppose the regression line of Y on X is,

$$Y = a1 + bxy\ X$$

$$byx = r\frac{\sigma y}{\sigma x} \quad \text{and}$$

$$= 0.8 \times \frac{3.0}{2.6}$$

$$= 0.923$$

and $a_1 = Y - b_1X = 12 - 0.923\ X5 = 7.385$

Thus, the regression equation of Y on X is

$$Y = 7.385 + 0.923\ X$$

In the like manner, we fit in the regression line of X on Y

$$bxy = r\frac{\sigma x}{\sigma y}$$

The regression line of X on Y is

$$X = -3.28 + 0.69\ Y$$

To Test

Ho : $\beta_{xy} = o$ vs. H1 : $\beta_{xy} \neq 0$

We compute the value t, firstly,

$$S_e^2 = \frac{1}{n-2}\left(\Sigma u_1^2 - bxy \quad \Sigma uivi\right)$$
$$= \frac{1}{7}\left(60.84 - 0.69 \times 6.24\right)$$
$$= 8.07$$

To test the significance of the regression coefficient, we establish the hypothesis as,

Ho : β_{yx} o vs H1 : $\beta_{yx} \neq o$

The residual variance,

$$S_e^2 = \frac{1}{7}\left(81 - 0.923 \times 6.24\right)$$
$$= \frac{1}{7}\left(81 - 5.76\right) = \frac{75.24}{7} = 10.75$$

∴ Again

$$Sb_{yx}^2 = \frac{10.75}{60.84}$$
$$= 0.177$$

Thus ,

$$t = \frac{0.923}{0.42}$$
$$= 2.14$$

Tabulated value of t for $\alpha \ll 0.05$ and 7 d.f is 2.365. Since the calculated value of t is less than the tabulated value of t, we accept Ho, it means, the regression coefficient of Y on X is nonsignificant.

To test the hypothesis,

$$Ho : a_0 = o \text{ vs. } H1 : \alpha \neq 0$$

We calculate Sa.

$$S_{e1}^2 = S_e^2\left\{\frac{1}{n} + \frac{\bar{x}^2}{\Sigma u_1^2}\right\}$$

$$= 10.75\left\{\frac{1}{9} + \frac{25}{60.84}\right\}$$

$$= 10.75 \times 0.522$$

$$= 5.6115$$

$$\therefore Sa = 2.37$$

The statistic value is

$$t = \frac{a1}{Sa1}$$

$$= \frac{7.385}{2.37}$$

$$= 3.11$$

$t_{0.05,7} = 2.365$. Since the calculated value $t = 3.11 > 2.365$, we reject H_0. It means the intercept a is significant at 5 per cent probability.

$$aS_{bxy}^2 = \frac{S_e^2}{\Sigma u_1^2}$$

$$= \frac{8.07}{81}$$

$$= 0.0996$$

$$Sbxy = 0.0996$$

Thus, the statistic

$$t = \frac{0.69}{0.0996}$$

$$= 6.093$$

Thus, the statistic

Again $t_{0.05,7} = 2.365$ which is less than 6.093. Hence, we reject Ho. Which means that the regression coefficient of X on Y is significant at 5 per cent probability.

$H_o : a_2 = 0$ vs. $H_1 : a_2 \neq 0$

$$S_{a2}^2 = S_e^2\left\{\frac{1}{n}+\frac{Y}{\Sigma u_1^2}\right\}$$

$$= 8.07\left\{\frac{1}{9}+\frac{144}{81}\right\}$$

$$= 8.07\{0.111+1.778\}$$

$$= 15.24$$

$$\therefore Sa = 3.90$$

$$t = \frac{a2}{sa2}$$

$$t = \frac{-3.28}{3.90}$$

$$= -0.84$$

The calculated t - value -0,84 is less than to0.5,7 =2.365, we accept Ho. It means that a_2 is nonsignificant.

(iv) Confidence interval for β_{Xy} is,

$$b_{xy} \mu t_a (n-2) S_{bxy}$$

We have already obtained

$b_{xy} = 0.69$ and $S_{bxy} = 0.996$ and $0.05, 7 = 2.365$

Substituting the values in the above formula, the confidence interval is,

$$0.69 \mu 2.365 \times 0.0996$$

$$0.69 \mu 0.236$$

The lower limit = 0.456

The upper limit = 0.926

Example-4: For the data given in example-1, find 95 per cent confidence limits for β_{yx}

Solution: Here we make use of the computations already made in example-1. The values are

$$B = 0.286 \text{ and } S_b = 0.0279$$

Also the tabulated value of t for a = 0.05 and 7 d.f. is 2.365. Thus 95 per cent confidence limits for by the formula (12) are,

$$C.L = 0.286 \mu 0.0279 \times 2.365$$

$$= 0.286 \mu 0.066$$

Thus,

95 per cent lower limit of = 0.220

95 per cent upper limit of = 0.946

Note: If we consider the regression of X on Y, all the formulae and procedures can be followed in the like manner simply by changing Y by X and X by Y.

QUESTIONS

1. Why should one test the significance of regression coefficient?

2. If We intercept comes out to fee non-significant, what do you infer by it?

3. How will you perform the test of significance for the regression coefficient?

4. Please give the formula for the standard error of b.

5. How will you test the significance of the intercept a in a regression line of Y on X?

6. What changes would you make in testing the significance of the regression coefficient of X on Y?

7. On the basis of the figures recorded below for supply and price for nine years, build a regression of price on supply

Also test the significance of the regression coefficient.

Supply :	80	82	86	91	83	85	89	96
93								
Price :	145	140	130	124	133	127	120	110
116								

8. Given the values for 10 paired observations on the variables X and Y as, $r = 0.6$, $\sigma x = 1.5$, $\sigma y = 2.0$, $\bar{X} = 10$ and $\bar{Y} = 20$.

(i) Calculate the regression lines of X on Y and Y on X.

(ii) Test the-significance of regression coefficients.

(iii) Test the significance of the intercept of Y on X.

(iv) Find 99 per cent confidence interval for β_{xy}

9. The sales and profit of a firm during the last 12 years were as follows:

Years :	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993
Sales :	62	74	56	50	48	32	52	53	41	47	60	44
(X:Lac.Rs.)												
Profit :	22	32	28	26	10	18	26	28	14	22	24	20
(Y: Lac Rs.)												

(i) Fit in regression line of Yon X.

(ii) Test the significance of regression coefficient β_{yx}

(iii) Find 95.per cent confidence interval for β_{yx}

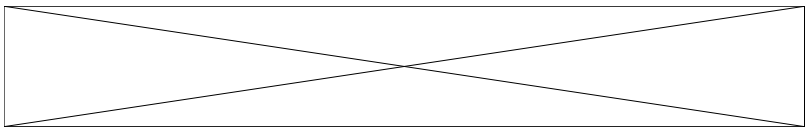
- End Of Chapter -

LESSON - 16

MULTIPLE LINEAR REGRESSION

Introduction

There are number situations where variable depends on, more than one independent variable In this situation, the estimated the of the dependent variable through single variable cannot yield satisfactory results For instance, the selling price of a finished product depends on the cost of the raw, material, labour cost, cost of energy (electricity), transportation, advertising cost etc. The production cost of a cereal depends on the cost of the feed, fertilizer, labour, irrigation etc. In all such situations to estimate the cost of "production can be estimated through a number of variables and hence one has to choose a multiple regression mode instead of a simple linear regression model.



Please use headphones

A multiple linear regression model with a dependent variable Y and K independent variables x_1 x_2,..... x_k can be given as,

$$Y= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon + \ldots\ldots + \beta_k x_k + \varepsilon \qquad \ldots\ldots\ldots\ldots \qquad (1)$$

β_0, β_1,..... β_k are regression constants and e is a random error term distributed normally with mean 0 and variance a σ_e^2 The variable Y is also distributed normally as $v(0,\sigma_y^2)$ Fitting of the above regression model means the estimation of the regression constants β_0, β_1,..... β_k that the error ε is minimized. The regression parameters can easily be estimated by theme; method of least squares. Instead of taking .the general model, we will confine out discussion taking k = 2 i.e. two variables.

Now we consider the linear regression model with two independent variables X_1 and X_2

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad \ldots\ldots\ldots\ldots\ldots \quad (2)$$

Suppose the above regression equation is to be fitted through n observational data consisting of the triplets as follows:

Y	X_1	X_2
Y_1	X_{11}	X_{21}
Y2	X12	X22
.	.	.
.	.	.
Y_i	Xl_i	$X2_i$
.	.	.
.	.	.
Y_n	Xl_n	$X2_n$

$$Y = b_0 + b_1 x_1 + b_2 x_2 + e \qquad \ldots\ldots\ldots\ldots\ldots \quad (2)$$

We know that all the triplets will satisfy the equation if they belong to it. Hence, for the i th triplet, the equation (2) can be written as

$$Y_i = \beta_o + \beta_1 \chi l_i + \beta 2 \chi 2_i + \varepsilon_i \qquad \ldots\ldots\ldots\ldots\ldots (3)$$

$$\varepsilon_i = (Y_i - \beta_o - \beta_1 \chi l_i - \beta 2 \chi 2_i)$$

Squaring both sides and taking sum over all observations, we get,

$$\sum_i \varepsilon_i^2 = \sum_i (Y_i - \beta_0 - \beta 1 \chi li - \beta 2 \chi 2i)^2$$

for I = 1,2,.....n

Let us put $\sum_i \varepsilon_i^2 = Q$

Thus $Q = \sum_i (Y_i - \beta_0 - \beta 1 \chi li - \beta 2 \chi 2i)^2$

Now we minimize Q by the method of least squares. Under this method differentiate Q partially with respect to β_o,β_1 and β_2 respectively and equate them to zero. Also replace β_o,β_1 and β_2 by b_o,b_1 and b_2 respectively, In this way we get three normal equations as follows:

$$\frac{\partial Q}{\partial \beta 0} = -2\sum_i (Yi - b0 - b1X\,1i - b2X\,2i) = 0 \qquad \text{.....(4)}$$

$$\frac{\partial Q}{\partial \beta 1} = -2\sum_i (Yi - b0 - b1X\,1i - b2X\,2i)X\,1i = 0 \qquad \text{.....(5)}$$

$$\frac{\partial Q}{\partial \beta 2} = -2\sum_i (Yi - b0 - b1X\,1i - b2X\,2i)X\,2i = 0 \qquad \text{.....(6)}$$

Rearranging the above equations, we get.

$$nb0 + b1\sum_i X\,1i + b2\sum_i X\,2i = \sum_i Yi \qquad \text{.....(7)}$$

$$b0\sum_i X\,1i + b1\sum_i X\,{}^2_1 i = b2\sum_i X\,2i = \sum_i X\,1i\;\;Yi \qquad \text{.....(8)}$$

$$b0\sum_i X\,2i + b1\sum_i X\,1i\;X\,2i = b2\sum_i X\,{}^2_2 i = \sum_i X\,2i\;\;Yi \qquad \text{.....(9)}$$

Rearranging the above equations, we get.

$$nb0 + b1\sum_i X\,1i + b2\sum_i X\,2i = \sum_i Yi \qquad \text{.....(7)}$$

$$b0\sum_i X\,1i + b1\sum_i X\,{}^2_1 i = b2\sum_i X\,2i = \sum_i X\,1i\;\;Yi \qquad \text{.....(8)}$$

$$b0\sum_i X\,2i + b1\sum_i X\,1i\;X\,2i = b2\sum_i X\,{}^2_2 i = \sum_i X\,2i\;\;Yi \qquad \text{.....(9)}$$

Equations (7), (8) and (9) are the normal equations.

Dividing equation (7) by n, we get

$$b0 + b1,\ \frac{1}{n}\sum_i X\,1i + b2,\ \frac{1}{n}\sum_i X\,2i = \frac{1}{n}\sum_i Yi$$

$$b0 + b1\,\bar{x}1 + b2\bar{x}2 = \bar{y}$$

$$b0 = \bar{y} - b1\bar{x}1 - b2\bar{x}2 \qquad \text{.....(10)}$$

Substituting the value of to in equations (8) and (9) from (10), we get.

$$\sum_i X\,1i(\bar{y} + b1\bar{x}1 - b2\bar{x}2) + b1\sum_i X_1^2 i + b2\Sigma i X\,1i\;X\,2i = \sum_i X\,1i\;yi$$

$$\sum_i X\,1i\;\;\bar{y} + b1\left(\sum_i X^2 - \sum_i X\,1i\;\bar{X}1\right) + b2\left(\sum_i X\,1i\;\;2i - \sum_i X\,1i\;\bar{X}2\right) = \sum_i X\;1i\;\;yi$$

$$b1\left(\sum_i X_1^2 i - n\bar{X}^2\right) + b2\;\Sigma i\,(X\,1i - \bar{X}1)(X2i - \bar{X}2) = \sum_i (X\,1i - \bar{y})(yi - \bar{y})$$

Now putting $\chi 1i - \bar{\chi}1 = u\,sini, \chi 2i - \bar{\chi}2 = u2i$ and $\bar{y}i - \bar{y} = vi$ we obtain

$$b1\Sigma u_{1i}^2 + b2\sum_i u\;1i\;u\;\;2i = \sum_i u\;1i\;\;yi$$

Similarly equation (9) can be reduced to

$$b1\sum_i u\;1i\;u\;2i + b2\sum_i u_2^2 i = \sum_i u\;2i\;ui$$

Now we have to solve the equations (11) and (12) for bi and b_2.

Method 1 : One way is to solve them by the method of elimination. Thus, in this way we get,

$$b1 \quad = \frac{\left(\sum_i uli\;\;vi\right)\left(\sum_i u_{2i}^2\right) - \left(\sum_i u\;1i\;\;u\;2i\right)\left(\sum_i u\;2i\;vi\right)}{\left(\sum_i u_{1i}^2\right)\left(\sum_i u_{2i}^2\right) - \left(\sum_i u\;1i\;u\;2i\right)^2}$$

$$b2 \quad = \frac{\left(\sum_i u2i\;\;vi\right)\left(\sum_i u_{1i}^2\right) - \left(\sum_i u\;1i\;\;u\;2i\right)\left(\sum_i u\;1i\;vi\right)}{\left(\sum_i u_{1i}^2\right)\left(\sum_i u_{2i}^2\right) - \left(\sum_i u\;1i\;u\;2i\right)^2}$$

Once we know b0, b1 and b2, the estimated regression equation is,

$$Y = Y + b1(\chi 1 - \bar{\chi}1) + b_2(\chi 2 - \bar{\chi}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(15)$$

Method 2 : Another simple way of solving the equations (11) and (12) could be through the determinants. In this approach we directly write the equation by taking unknowns in the numerator and determinants in the denominator leaving its known coefficients in a particular order as given below :

$$\frac{b1}{\begin{vmatrix} \Sigma u\ 1i\ vi & \Sigma u\ 1i\ v2i \\ \Sigma u\ 2i\ vi & \Sigma u\ {}_2^2 i \end{vmatrix}} = \frac{-b2}{\begin{vmatrix} \Sigma u\ 1i\ vi & \Sigma u\ {}_1^2 i \\ \Sigma u\ 2i\ vi & \Sigma u\ 1i\ v2i \end{vmatrix}} = \frac{1}{\begin{vmatrix} \Sigma u\ {}_1^2 i & \Sigma u\ 1i\ v2i \\ \Sigma u\ 1i\ v2i & \Sigma u\ {}_2^2 i \end{vmatrix}}$$

From the above equations we obtain, and

$$b1 = \frac{\left(\sum uli\ vi\right)\left(\sum u_{2i}^2\right) - \left(\sum u\ 1i\ u\ 2i\right)\left(\sum u\ 2i\ vi\right)}{\left(\sum u_{1i}^2\right)\left(\sum u_{2i}^2\right) - \left(\sum u\ 1i\ u\ 2i\right)^2}$$

$$b2 = \frac{\left(\sum u2i\ vi\right)\left(\sum u_{1i}^2\right) - \left(\sum u\ 1i\ u\ 2i\right)\left(\sum u\ 1i\ vi\right)}{\left(\sum u_{1i}^2\right)\left(\sum u_{2i}^2\right) - \left(\sum u\ 1i\ u\ 2i\right)^2}$$

Since $[A] = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = (a_{11}a_{22} - a_{12}a_{21})$ b_1 and b_2 from (16) (17) are same as in (13) and (14). Regression equation (15) stands as such.

Method 3: Earlier two methods are simple and applicable for two unknowns: But the modern approach is to solve the equation through matrix approach.

We know, the equations (11) and (12) in the matrix notations can be written in the following manner.

$$\begin{vmatrix} \Sigma u\ {}_1^2 i & \Sigma u\ 1i\ v2i \\ \Sigma u\ 1i\ v2i & \Sigma u\ {}_2^2 i \end{vmatrix} \begin{bmatrix} b1 \\ b2 \end{bmatrix} = \begin{vmatrix} \Sigma u\ 1i\ vi \\ \Sigma u\ 2i\ vi \end{vmatrix} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(18)$$

If we put,

$$A\,2\times 2 = \begin{vmatrix} \Sigma u\ {}_1^2 i & \Sigma u\ 1i\ v2i \\ \Sigma u\ 1i\ v2i & \Sigma u\ {}_2^2 i \end{vmatrix} \quad B\,2\times 1 = \begin{bmatrix} b1 \\ b2 \end{bmatrix} \quad \text{and}$$

$$V\,2\times 1 = \begin{vmatrix} \Sigma u\ 1i\ vi \\ \Sigma u\ 2i\ vi \end{vmatrix}$$

Matrix equation (18) is equivalent to

$$A\,B = v \qquad \ldots\ldots\ldots\ldots\ldots\ldots(19)$$

or $B = A^{-1}\,v \qquad \ldots\ldots\ldots\ldots\ldots\ldots(20)$

Inverse matrix A^{-1} can be obtained by the formula.

$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} A11 & A12 \\ A21 & A22 \end{bmatrix}$$

Where $\begin{bmatrix} A11 & A12 \\ A21 & A22 \end{bmatrix}$ is an adjo int matrix.

Thus

$$[A] = (\Sigma u_{1i}^2)(\Sigma u_{2i}^2) - (\Sigma u1i\ u2i) = D(say)$$

and

$$A11 = \Sigma u_{1i}^2, A12 = A21 = -\Sigma u1i\ u2i,\ A22 = \Sigma u_{1i}^2$$

Equation (20) can be written as,

$$\begin{bmatrix} b1 \\ b2 \end{bmatrix} = \begin{bmatrix} \frac{\Sigma u_1\ {}_2^2 i}{D} & \frac{\Sigma u_1\ 1i\ u2i}{D} \\ \frac{-\Sigma u\ 1i\ u2i}{D} & \frac{\Sigma u_1\ {}_1^2 i}{D} \end{bmatrix} \begin{bmatrix} \Sigma u\ 1i\ vi \\ \Sigma u_1\ 2i\ vi \end{bmatrix}$$

By equivalence

$$b1 = \frac{\left(\sum u1i\ \ vi\right)\left(\sum u_{2i}^2\right) - \left(\sum u\ 1i\ \ u\ 2i\right)\left(\sum u\ 2i\ vi\right)}{D}$$

$$b2 = \frac{\left(\sum u2i\ \ vi\right)\left(\sum u_{1i}^2\right) - \left(\sum u\ 1i\ \ u\ 2i\right)\left(\sum u\ 1i\ vi\right)}{D}$$

Again we get the same results.

Note: It does not matter what way we$_t$ calculate b_1 and b_2. We get the same linear regression equation.

Partial regression coefficient:

Any of the coefficient β_1 or β_2 of $\chi_1$or χ_2 is called the partial regression coefficient. β_1, is more. Specifically denoted as$\beta_{y1.2}$ which is self explanatory that it is the partial regression coefficient of Y on χ_1 excluding χ_2 Similarly β_2 can be denoted as $\beta_{y2.1}$ Their estimates are denoted as $6_{y1.2}$. and $b_{y2.1}$ respectively. But usually the suffixes are not attached in full. And are understood by himself.

Definition:

The partial regression the coefficient β_j (j=1,2), the coefficient of χ_I is a measure change in the dependent variable Y corresponding to an unit change in the variable x i eliminating the effect of the other variable (s).

Test of significance of Partial regression coefficient

Whatever may be the estimated value of the partial regression coefficient, one wants. To test its significance as it testifies whether the contribution by a particular independent variable x 2 in estimating Y is of relevance or not. Here we test the hypothesis

$H_0 : \beta_j = 0$ us $H_1 : \beta j \neq 0$

The hypothesis H_0 can be tested by t - test where the statistic

$$t = \frac{bj}{Sbj} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (24)$$

t less (n-3) d.f for j = 1,2.

Where b_j is the estimated value of β_j and sbj is standard error of b_j

S_{bj} can be calculated in the following manner. First calculate $S\&$ by the formula.

$$S_E^2 = \frac{\sum_i (Yi - Yi)^2}{n-3} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (25)$$

$$= \frac{1}{n-3}\left(\sum_i u_1^2 - R^2 \Sigma V_1^2\right) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (26)$$

Where

$$R^2 \Sigma u_1^2 = b1 \Sigma u1ivi + b2 \sum_i u2ivi \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (27)$$

The quality $R^2\Sigma u_1^2$ is known as the regression sum of squiarres. Also the ratio $R^2 = \frac{R^2\Sigma u_1^2}{\Sigma u_1^2}$ is called the coefficient of det er min ation. It is the positive square root R is called the multiple corrlation. A higher value means a value more than 0.9. Also the quantity $(1-R^2)$ is known as the coefficient of alientatio n. We know all terms in (27). Hence $R^2\Sigma u_1^2$ is easy to calculate.

$$S_{bj}^2 = S_E^2 . Cjj \qquad \ldots\ldots\ldots\ldots\ldots\ldots (29)$$

Where $Cjj = \Sigma u_{jj}^2 / D$ for $j = 1,2.$

Decision about*Ho is easily taken by comparing calculated value of- t with the tabulated value of t for a level of significance and- (n -. 3) degrees of freedom. If $t_{cal} > t_\alpha$, $(n - 3)$. reject H_o. It means the partial regression coefficient β_j is significant, otherwise reject Ho.

Confidence interval for β_j

Suppose one wants (1 - ***α)*** 100 per cent confidence interval for the partial regression coefficient. In the usual way, the confidence internal for β_j can be given by the formula.

$$bj \pm ta,(n-3)Sbj \qquad \ldots\ldots\ldots\ldots\ldots(30)$$

We already know b_j and S_{bj} t_a (n - 3) .is the Calculated value oft for α level of significance and (n — 3) degrees of freedom.

Example - 1 The data with regard to the out put of gram and the cost of seed and labour per hectare at ten farmer's fields are given below :

S.No.	Cost of Produce (Y) Rs./ha	Cost of Seed (X_1) Rs. / ha	Labour cost (X_2) Rs. / ha
1.	1127	235	128
2.	840	236	82
3.	735	234	204
4.	570	241	71
5.	463	238	110
6.	614	233	130
7.	916	235	200
8.	460	190	170
9.	1540	235	180
10.	1065	243	165

(i) Fit the regression equation $\hat{y} = b_0 + b_1\chi + b_2\chi_2$

(ii) Estimated the cost of product per hectare given that $\chi 1$ = 230 and $\chi 2$ = 125

(iii) Test the significance of partial regression coefficients

(iv) Find the 95 per cent confidence interval for b_1

Solution

To fit in the regression equation we prepare the following calculation table

S.No	Y	XI	X2	Y-T υ=	X_1-X_2	X_2-X_2 = u_2	u_1u_2	u1v	$u2v$	u_1^2	u_2^2	v^2
1.	1127	235	128	294	3	- 16	- 48	882	-4704	9	256	86436
2.	840	236	82	7	4	- 62	- 248	28	- 434	16	3844	49
3.	735	234	204	- 98	2	60	120	- 196	- 5880	4	3600	9604
4.	570	241	71	- 263	9	- 73	- 657	- 2367	19199	81	5329	69169
5.	463	238	110	- 370	6	- 34	- 204	- 2220	12580	36	1156	136900
6.	614	233	130	- 219	1	- 14	- 14	-219	3066	1	196	47961
7.	916	235	200	83	3	56	168	249	4648	9	3136	6889
8.	460	190	170	- 373	- 42	26	- 1092	15666	- 9698	1764	676	139129
9.	1540	235	180	707	3	36	108	2121	25452	9	1296	499849
10.	1065	243	165	232	11	21	231	2552	4872	121	441	53824
Total	8330	2320	1440	0	0	0	- 1636	16496	49101	2050	19930	1049810

(i) Making use of the computations done in the above table, we find all required values.

$Y = \frac{8330}{10} = 833,\ X1 = \frac{2320}{10} = 232$ and $X2 = \frac{1440}{10} = 144$

The quantity D in (22) and (23) is,

$$D = (2050)(19930) - (-1636)^2$$
$$= 40856500 - 2976496$$
$$= 38180004$$

$$b1 = \frac{(16496)(19930) - (-1636)(49101)}{38180004}$$
$$= \frac{328765280 + 80329236}{38180004}$$
$$= \frac{409094516}{38180004}$$
$$= 10.71$$

$$b2 = \frac{(49101)(2050) - (-1636)(16496)}{38180004}$$

$$= \frac{100657050 + 26987456}{38180004}$$

$$= \frac{127644506}{38180004}$$

$$= 3.34$$

Thus the required regression equation is

$$Y = 833 + 10.71 (\chi_1 - 232) + 3.34 (\chi_2 - 144)$$

$$Y = -2132.68 + 10.71\ \chi_1 + 3.34\ \chi_2$$

(ii) The estimated value of Y for $\chi_1 = 230$ and $\chi_2 = 125$ is

$$Y = -2132.68 + 10.71 \times 230 + 3.34 \times 125'$$

$$= 748.12$$

(iii) For performing the test of significance of the regression coefficients, first we calculate S^2_e from (27).

$$R^2\Sigma u_1^2 = 10.71 \times 16469 + 3.34 \times 49101$$
$$= 176382.99 + 163997.34$$
$$= 340380.33$$
$$S_e^2 = \frac{1}{(10-3)}[1049810 - 340380.33]$$
$$= \frac{709429.67}{7}$$
$$= 101347.10$$
$$C11 = \frac{2050}{38180004}$$
$$= .000522$$
$$S_{b1}^2 = 101347.1 \times 0.0000537$$
$$= 5.44$$
$$S_{b1} = 2.33$$
$$S_{b2}^2 = 101347.1 \times .000522$$
$$= 52.90$$
$$S_{b2} = 7.27$$

To test,

$$H_0 : \beta 1 = 0 \text{ us. } H_1 : \beta \neq 0$$
$$t = \frac{10.71}{2.33}$$
$$= 4.60$$

Tabulated value oft for $\alpha = 0.05$ and 7 d.f. is 2.365. Since the calculated t value 4.60 is greater than the tabulated value t 2.365, we reject Ho. Hence* we conclude that the partial regression coefficient *Or* is significant. Again to test,

Again to test,

$$H_0 : \beta 2 = 0 \text{ us. } H1 : \beta \neq 0$$
$$t = \frac{3.34}{7.27}$$
$$= 0.46$$

Calculated value of t = 0.46 is less than t 0.05,**7** = 2.365, we accept Ho. It means that the partial regression coefficients β_2 is non - significant:

(iv) 95 per cent -confidence- interval for b_1 is given by the formula

$b1 \pm S_{b1,}\ t_{0.05},7$

10.71 ±2.33 x 2.365

10.71 ±5.51

The Lower limit = 520

The upper limit = 16.22

Addendum :

Here it is to add further that the equations (7) through (9) can be solved to obtain b_0, b_1 and $b_2$wihtout taking: the, deviations from respective means. The matrix approach is the best. The readers are referred to 'Basic Statistics' by B.L. AGARWAL, Chapter 14.

The three normal equations in matrix notations can be written as.

$$\begin{bmatrix} n & \sum_i X_{1i} & \sum_i X_{2i} \\ \sum_i X_{1i} & \sum_i X_{1i}^2 & \sum_i X_{1i} X_{2i} \\ \sum_i X_{2i} & \sum_i X_{1i} X_{2i} & \sum_i X_{2i}^2 \end{bmatrix} \begin{bmatrix} b0 \\ b1 \\ b2 \end{bmatrix} = \begin{bmatrix} \sum_i Y_i \\ \sum_i X_{1i} Y_i \\ \sum_i X_{2i} Y_i \end{bmatrix}$$

Suppose, the matrices

$$\begin{bmatrix} n & \sum_i X_{1i} & \sum_i X_{2i} \\ \sum_i X_{1i} & \sum_i X_{1i}^2 & \sum_i X_{1i} X_{2i} \\ \sum_i X_{2i} & \sum_i X_{1i} X_{2i} & \sum_i X_{2i}^2 \end{bmatrix} = A$$

$$\begin{bmatrix} b0 \\ b1 \\ b2 \end{bmatrix} = B \quad \text{and} \quad \begin{bmatrix} \sum_i Yi \\ \sum_i X_{1i}\, Yi \\ \sum_i X_{2i}\, Yi \end{bmatrix} = Y$$

The matrix A is also known as coefficient matrix.

The normal equations are,

$$A\,B = Y$$

or $$B = A^{-1}\,Y$$

The inverse of A can obtained by pivotal condensation method, see appendix A of Basic Statistics by B L. Agarwal.

Let the inverset matrix of A is

$$A^{-1} = \begin{bmatrix} C01 & C02 & C03 \\ C11 & C12 & C13 \\ C21 & C22 & C23 \end{bmatrix}$$

Matrix A is symmetrical matrix and A $^{-1}$ will also be a symmetrix matix. Once we know A^{-1}, we can write (33) in the expanded form as,

$$\begin{bmatrix} b0 \\ b1 \\ b2 \end{bmatrix} = \begin{bmatrix} C01 & C02 & C03 \\ C11 & C12 & C13 \\ C21 & C22 & C23 \end{bmatrix} \begin{bmatrix} \sum_I Yi \\ \sum_I X_{1i}\, Yi \\ \sum_i X_{1i}\, Yi \end{bmatrix}$$

From (34);

$$b0 = C01 \sum_i Yi + C02 \sum_i X_{1i}\, Yi + C03 \sum_i X_{2i}\, Yi$$

$$b1 = C11 \sum_i Yi + C12 \sum_i X_{1i}\, Yi + C13 \sum_i X_{2i}\, Yi$$

$$b3 = C21 \sum_i Yi + C22 \sum_i X_{1i}\, Yi + C23 \sum_i X_{2i}\, Yi$$

Once we know b_0, b_1, and b2, the regression equation can easily be written. Further treatment can be extended by symmetry.

Multicollinearity:

The term multicollinearity is used to describe a situation where the explanatory variables X_1,X_2,X…….X_k a Ere correlated. If multicollinearity is present, the regression model-will riot-yield good results'[1] The problem of multicollinearity is usually confronted with in time series data.

If perfect multcollinearily Is present, the coefficient matrix is .usually singular i.e. the determinant of the coefficient matrix is zero. Consider the case of 4 regressor variables X_1 and X_2. If X_1 and X2 are correlateds $X_1 = CA$ &.The coefficient matrix,

$$A = \begin{bmatrix} \Sigma \chi_{1i} & \Sigma \chi_{1i} \chi_{2i} \\ \Sigma \chi_{1i} \chi_{2i} & \Sigma \chi_{2i} \end{bmatrix} = \begin{bmatrix} C^2 \Sigma \chi_{2i}^2 & C \Sigma \chi_{2i}^2 \\ C \Sigma \chi_{2i}^2 & \Sigma \chi_{2i}^2 \end{bmatrix}$$

$$\therefore |A| = 0.$$

The above situation exists in case of perfect collinearity which in practical life rarely exist. What we usually come across is the high collinearity.

In case of high collinearity, the best remedy is to redefine the mix of the variables by either discarding or combining some of the variables. Usually dropping of a variable introduces b_1 as in the regression coefficient. and combining one or two variables, one looses information. -

A remedy to multicollinearity depends on the order of collinearity, purpose of analysis of data, the relative importance of the variables included etc.

For details on multicollinearity, the readers are referred to

(i) Introduction to, Econometrics, By L.R. Klein -

(ii) Applied multivariate data analysis, by J.D. Jobson

QUESTIONS

1. What is the importance of multiple regression analysis ?
2. Discuss a statistical multiple regression model ?
3. How do you fit in a multiple regression equation?
4. Define partial regression coefficient.
5. How can one adjudge the appropriateness of linear model ?

6. The daily rainfall of 14 selected places along with the altitude and distance from the sea level was as follows:

Place No.	Rainfall (mm)	Altitude (I.e.)	Distance (Km.)
1.	161.0	156	330
2.	173.4	268	306
3.	201.7	358	492
4.	229.0	451	564
5.	248.4	653	708
6.	201.8	443	600
7.	247.6	'593	666
8,	229.9	359	528
9.	226.6	313	516
10.	216.2	422	618
11.	218.5	229	444
12.	207.1	619	612
13.	181.3	370	636
14.	189.7	685	648

(i) Fit in the linear regression of rainfall on altitude and distance.

(ii) Test the significance of partial regression coefficients.

(iii) Estimate the rainfall for given altitude = 600 and distance = 600.

(iv) Establish 99 per cent confidence limits for the partial regression coefficients

7. The following table gives the number of leaves per plant, height of plant and height of main stem of mung at eleven places.

No. of leaves per plant (Y)	Height of Plant (cms) (Xi)	Height of main stem (cms) (X_2)
21.0	45.5	24.9
21.2	52.0	32.2
22.6	59.1	40.9
21.6	57.4	38.4
21.7	56.8	37.1
22.2	59.1	40.6
21.6	47.5	27.2
226	59.5	40.7
22.4	58.4	39.4
22.8	61.7	42.2
23.4	58.5	38.8

(i) Fit in the regression equation $\hat{y} == b_0 + b_1 X_1 + b_2 X_2$

(ii) Test the significance of β_2.

(iii) Find 95% confidence limits for β_1

(iv) Estimate the number of leaves per plant, for $X_1 = 55$ and $X_2 = 36$

8. How can you test the significance of a partial regression, coefficient ?

9. How to find out the confidence limits of a partial regression coefficient ?

10. What do you understand by multi collinearity ?

11. How multi collinearity effects the regression coefficient ?

12. How can one tackle with the problem of multi collinearity ?

- End Of Chapter -

LESSON - 17

STATISTICAL DECISION THEORY

Preamble

There is hardly any moment in life left without decision making. One starts taking decision right from morning whether to get up in the early morning or to continue sleeping till late hours, whether to buy an article or not. Even if to buy an article, then from which shop he should buy it; what quality product one should buy it. All such decisions are common decision and are based purely on the judgment, experience, liking and will of an individual.

In scientific phenomenon, testing of hypothesis comes under the category of classical decision theory. In this case, we have an assertion about population parameters and has to decide whether to accept or reject the hypothesis on the basis of certain appropriate statistical test at certain level of significance. Classical decision theory suffers with three aspects.

Firstly it takes only two actions about the null hypothesis vis-a-vis alternative hypothesis whether to accept or reject the null hypothesis.

Secondly it does not take into account any other information pertaining to the decision except empirical data which is being collected through the sampling process.

Thirdly, there are economic consequences that result from making a wrong decision. Although such consequences are covered by taking into consideration a desired level of significance level for the test. These procedures are never the explicit part of the decision model or procedure. But some prior information is utilized under Bayesian decision theory as this is based upon a direct evaluation of the payoff for such alternative course of action.

The Bayesian decision theory or simply decision theory removes the above short-comings and enables are to take optimal decisions. The reasons for optimal decision being that:

(i) It provides a model for decision making in situations that involves multiple states of the parameter which is termed as nature-in parlance in decision theory

(ii) It incorporates the economic consequences of making a wrong decision.

(iii) It utilises the information .pertaining to the decision making which exists prior to any sampling or experiment. The prior information may be in the form of empirical data or a subjective type of information considered useful by the decision maker. For instance, one has to undergo an operation. He takes into consideration the survival rate of patients after the operation and also the expertise of the doctor. Survival rate is an empirical data whereas expertise of the doctor is a subjective consideration.

Decision making is extremely useful in business. A wrong decision puts the business organisation into heavy losses and even the company or business may fail. How much stocks should be maintained,, what percentage of profit be fixed, what short of items be manufactured are the parts of decision making in business.

The likelihood principle: It is another principle which is largely used in decision making. This involves a likelihood function which is defined as, the function L(q,x), where the sample x has been observed,. This is considered as a function o£0 and is called the likelihood function. The function L(q,x), L(x/q).

The intuitive reason for the name likelihood function is that a θ for which f(x/q). is large is more likely will be true than a small value of f(x/q).

The likelihood principle:

In making decisions about when $x_;$ is observed all relevant information is contained, in. f(x / θ)Hence, this principle bears, good importance. The details are omitted as this does not constitute the part of the syllabus.

COMPONENTS OF THE DECISION PROBLEMS

The business organisation has to make decisions every day with regard to its expansion, number of units to be produced, price fixing, whether to replace the old plant with the. new one etc. All this has to be decided on some criteria. Therefore, various steps in consideration are called ingredients of the decision problem which are discussed in brief below.

1. Alternative course of action:

The decision problem arises only when there are different course of action at our disposal and a decision maker has to choose one out of many. Let there are K actions, a1, a2, a3,.....ak at the disposal of the decision makers and he has to choose one put of these alternative actions. The set of K actions is known as action space.

If the action selected does not fulfil the objective, it would result into the waste of time and cause heavy losses. So, by making use of all the available information, an action has to be chosen; based on statistical decision procedures which makes one to attain an optimal decision which fulfils the objectives i.e. which minimizes the loss and/or maximises the gain etc.

2. Uncertainty:

It is not possible to predict the outcome of an experiment. Hence, the outcome is said to be uncertain. So, there are many outcomes for an event which are called states of nature in decision theory. So, it is possible to predict the state of nature in terms of probabilities. In an usual manner K states of nature are represented by q_1 q_2 , . . . q_k .The, totality of states of nature; is called states space and is denoted by Q. If an action leads to four outcomes θ1,θ2 , θ3and θ4 Then W = (q_1, q_2, q_3, q_4)

For instance, a product liked by 100 per cent customers is denoted by q_1 , a product liked by 50 per cent person is denoted by q_2 ; the product liked by 25 percent buyers is denoted by q_3 and the product not liked by any is denoted by q_4.

As a matter of fact the decision making under 'risk' and decision making under uncertainty are not synonyms. They are different in the, sense that when the state of nature is unknown but objective or empirical data is available to enable are to assign probabilities to the various, states of nature, the procedure is referred to as decision making under 'risk'. But if the state of nature is unknown and there is no object or empirical data available to assign probabilities to various states of nature, then the decision procedure is referred to as decision under uncertainty. Anyhow, if the probabilities are assigned even sunder uncertainty on intuitive basis, the decision procedures under risk and under uncertainty are equivalent.

3. Pay Offs :

Usually one evaluates the consequences of a course of action for each event in terms of monetary value or time. A number of consequences result from each action under different states of nature. For ‘m’ possible acts and 'n' states of nature, there will be m x n consequences. The consequences are usually evaluated in monetary terms such as,

i. in terms of profit
ii. in terms of cost
iii. in terms of opportunity loss
iv. unit of utility

The payoff table can be presented in a two way tables as follows:

Table 1-1 Pay off table

States of	a_1	a_2	a_3	....	a_j	....	a_m
θ_1	P_{11}	P_{12}	P_{13}	....	$P1_j$	....	$P1_m$
θ_2	P_{21}	P_{22}	P_{23}	....	$P2_j$	....	$P2_m$
θ_3	P_{31}	P_{32}	P_{33}	....	$P3_j$	....	$P3_m$
.							
.							
.							
θ_i	p_{i1}	P_i2	P_i3	....	P_{ij}	....	Pi_m
.							
.							
.		P_n2	P_n3	....	P_{ij}	....	p_{nm}
θ_n	P_n1						

In table 1.1, P_{ij} represent the pay off as a consequence of act aj when the 'n' state of nature is q_i for i = 1,2,....n and j = 1,2,....m. With the help of the payoff table, a decision maker can reach an optimum solution of the problem in respect of an event which is going to occur. Since, the outcome which, is going to occur is unknown, a forecast has to be made in terms of probabilities assigned to the events to occur. The last, step is to make use of these probabilities for calculating the expected pay off of expected monetary value for each course of action. A decision maker has to choose an optimal act which results into maximum expected pay off.

An alternative way to decide about and act for the events is on the basis of expected opportunity loss (EOL). This criterion yields the same results which are obtained by expected profits.

Since the pay offs depend on the choice of a particular act and are conditional and subject to the occurrence of an event, are often termed as conditional values and the pay off table is termed as the conditional value table.

The payoff for any cell of the payoff table 1.1 calculated by the formula

Payoff = Demand x Sales price - Stock x Cost

Expected monetary value

Expected monetary value (EMV) is also called expected payoff. Suppose Xi denotes the ith event in an act and Pi is the probability that Xi takes place, the expected monetary value of an act is given as,

$$EMV = \Sigma \chi_i p_i$$

for = 1,2,3....

To make a decision for an act, the rule is to select an act which has maximum expected monetary value. The criterion of selecting the maximum EMV of an act is often referred to as Bayes decision rule named after Thomas Bayes.

Risk function

If q is the state of nature and 'a' is an act, the function R(q , a) is known as the risk function and denotes the risk involved in taking a decision about when the act 'a' has been adopted. The risk R(q, a) is nothing but the expected loss incurred in taking the act 'a' about q Hence,

$$R(\theta,a) = E[L(\theta,a)]$$

Loss is usually taken to mean opportunity loss (OL) of money, time, fuel etc. The loss function is a pessimistic view taken by the statisticians. Contrary to this, economists take an optimistic view and talk of utility function. It amounts to the same whether minimises the loss function or maximises the utility function.

Inadmissible act

An act is called an inadmissible act, if it is dominated by any other act. An inadmissible act in a payoff table is one for which the payoffs for an act are less than the payoffs for any other act. Corresponding to the events in this situation, the inadmissible act is discarded from the payoff table. This process of elimination $_{i;}$ saves labour of calculations and simplifies the process of decision making.

Example-1.1: Consider the following payoff table with five acts and four events.

Events	A_1	A_2	Acts A_3	A_4	A_5
E_1	6	8	15	6	12
E_2	7	5	2	13	10
E_3	9	4	8	9	3
E_4	12	11	18	15	14

From the above table it is apparent that the payoffs for the act A2 for all events are less than the payoffs for the Act A3. Therefore, the act A2 is inadmissible. This act can be removed from the table. In this way the payoff table reduces to the order of 4 x 4 for further analysis.

Expected opportunity loss

The name itself indicates that the loss incurred due to missing of better opportunity is termed as expected opportunity loss (EOL). Expected opportunity loss is an alternative to EMV approach. Thus, the expected opportunity for an outcome is the difference between the best pay off for an event and the payoff for the outcome of that event under an act. In other words it is the loss incurred due to the gain missed which could be earned by making the right choice of an act for an event. This has been shown in the following example.

Example-1.2 : Suppose a shopkeeper costs a particular type of sweet for Rs.5 each and sells it for Rs.6 each. Now the problem before the shopkeeper is that any sweets left unsold will be a net loss as it is a perishable item. So how much should he prepare. If the shopkeeper expects the sale of 100 items per day with probability 0.5, 150 items with probability 0.4 and 200 items with probability 0.1, then how much should be prepared so that the expected opportunity loss is minimum. The loss table can be prepared and displayed in the following manner.

Events	Probabilty	Acts Prepare	Prepare 150	Prepare 200
Sale of 100 items	0.5	0	250	500
Sale of 150 items	0.4	50	0	250
Sale of 200 items	0.1	100	50	0

In the above table, the loss for the act, prepare 100 and sale of 100 item is zero, as the shopkeeper prepared and sold 11 of them. So there is no loss of opportunity, whereas prepare 150 and only sold 100 item, 50 items are left unsold. In this way he has incurred a loss of Rs.250 directly.

Again if the shopkeeper has prepared only 100 items whereas he could sell 150 items, he has lost the opportunity of the profit on 50 items. So he has incurred a opportunity loss of Rs.50. The other entries are made likewise. Now the expected opportunity loss for each act can be calculated by taking the sum of the product of losses with their corresponding probabilities.

EOL for the Act, prepare 100 is,

EOL = 0. x 0.5 + 50 x .4 + 100 x .1

= 0 + 20 + 10 = 30

EOL for the Act prepare 150 is

EOL = 250 x 0.5 + 0 x 0.4 + 50 x 0.1

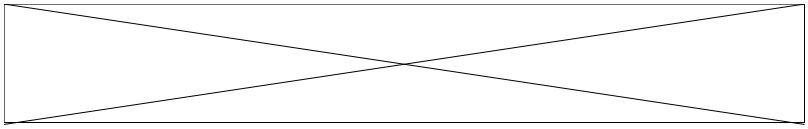
= 125 + 0 + 5= 130

EOL for the Act prepare 200 is

EOL = 500 x 0.5 + 250 x 0.4 + 0 x .1

= 250 + 100 + 0 = 350

Since the minimum expected opportunity loss is Rs.30 for the act-prepare 100 items, the shopkeeper should prepare 100 items.



Please use headphones

Optimal decisions

Decision theory helps to select an action which minimises the loss or maximises the gain. As we know, there is no single act which can always be best under all situations (states of nature). An act may be the best for particular state of nature and the worst of the other state of nature. So a decision maker has to choose one under uncertainties. Hence, a decision maker wants some criterion on the basis of which he can choose the best act out of the many at his disposal. Here, we shall discuss three principle which are popularly used, named:

(i) the Maximin principle

(ii) the Minimax principle

(iii) the Baye's principle.

Maximin principle

It is the simplest principle out of all principles for cheesing an optimal action when the payoffs are given in terms of profits. According to maximin principle a decision maker first selects the-minimum payoffs over the various possible states of nature. Then he "selects that action for which the minimum payoff.' This principle guards against worst that can happen and makes him prepared to face the worst. In short under maximin principle a decision maker chooses an act which maximises the min P_{ij}

Example-1.3 : For the problem given in example-1.2, we prepare the following payoff table

States of nature	Prepare – 100 (A_1)	Acts prepare – 150(A_2)	Prepare – 200 (A_3)
Demand - 100	100	-150	-400
Demand – 150	400	150	-250
Demand – 200	700	450	200

Pay off = Demand x Sales price – Stock x cost

P_{11} = 100 x 6- 100 x 5 = 100

P_{12} = 100 x 6- 150 x 5 = -150

P_{13} = 100 x 6- 200 x 5 = -400

Similarly

P_{21} = 400, P_{22} = 150, P_{23} = -250

P_{31} = 700, P_{32} = 450, P_{33} = 200

In the above payoff table, the act A1 has minimum payoff 100, act A2 has minimum payoff-150 and for act A3, the minimum payoff is -400. Out of the three minimum payoffs 100, -150, -400, the payoff100 is maximum and the shopkeeper should choose to prepare 100 items. This is the same decision which we obtained on the basis of expected opportunity loss.

Minimax principle

This principle is used when the payoffs are-given in terms of opportunity losses. Here we minimize the maximum opportunity joss. A decision maker first observes the maximum opportunity loss over all states of nature. He then chooses the action for which the maximum opportunity loss is minimum. This principle guards against the maximum loss.

Example-1.4 :

Now we consider the decision problem given in example-1.2 and, wish to select the best act on. the basis of minimax principle. From the loss table the maximum loss, for the act A_1 is 100, for apt A2 is 250 and for the act A_3 is 500. The minimum loss out of the maximum losses out of three acts is Rs. 100 for the act A_1. Hence, one should choose the act Ai again the same result is obtained as we got from maximin principle.

Baye's principle

Baye's principle is based on the use of a priori probabilities. A major advantage of Bayesion approach is that a decision maker selects an action on a rational basis since he uses subjective probabilities ascertained from his own experience, past performance or intuitively. Here we explain Bayesion principle of decision making in brief.

To make use of Bayesion principle, a decision maker should first assign prior probabilities to each of the state of nature. Here, it should be bear in mind that the sum of these probabilities is equal to 1. These probabilities reflect on the belief of the decision maker about the states of nature to occur. The set of probabilities along with the states of nature constitute a probability distribution known as prior distribution. Let the probability density function (Pd.f.) of the prior distribution for the state of nature *0* is g (q). For the state of nature$\theta = q_i$ the p.d.f. is g(q, i), Let us assume that L(q , g) is the loss function, which is non-negative.

Suppose X1, X2, ... X_n are continuous random variables having the joint probability density function f_θ (x). The risk function for a decision rule 'δ' such that $\delta \in 4$ (a set of decision rules) is given as,

$$r(g(\theta), x) = \int [1\theta L(\theta, \delta), gx(\theta)\, d\theta] g(x)\, dx \qquad (1.4)$$

Where

$$fg(x) = \int \theta\, f\theta(x)\, g(\theta)\, d\theta$$

and f_g (x) > θ

and

$$g_x(\theta) = \frac{f\theta(x) g(\theta)}{\int \theta\, f\theta(x) g(\theta)\, d\theta} \qquad (1.5)$$

$$= \frac{f\theta(x) g(\theta)}{fg(x)} \qquad (1.5.1)$$

It is commonly called the probability density function of the posterior distribution, of *0* , given that x has been observed. Having the idea about the prior and posterior distribution, we directly return to the decision problem.

Baye's approach states that the expected payoff for each act in a set of acts be computed and on the basis of these payoffs, a best act should be chosen A best act in case of profit is one for which the expected profit is maximum or the expected loss is minimum. This is also known as expected monetary value (EMV) criterion. A best act is one for which EMV is maximum. The procedure for this is-

i. Prepare a payoff table
ii. Assign probabilities to each of the event or state of nature.
iii. Calculate EMV for each act separately as the sum of the product of the payoff value and
iv. Probabilities assigned to each event.
v. Choose the act which has the maximum EMV

Once the prior distribution is determine, Baye's principle is applied to make a decision about selecting the optimal action. The Bayesion analysis is carried out in three phases namely, (i) Prior analysis (ii) preposterior analysis (iii) posterior analysis each of these are discussed in brief.

i. **Prior analysis** : Once a decision ma has determined the prior distribution- $g(\theta j)$ for the states of nature θj he expected payoff (EP) or expected opportunity loss (EOL) has to be worked out for each action aj by the formula,

(Expected payoff) $B(a_i) = \Sigma \; P_{ij} \; g(\theta_j)$ for all j

Instead of payoff if one uses opportunity losses, he gets EOL corresponding to action aj. The decision maker chooses an action for which the EP Is maximum or EOL is minimum.

Example-1.5 : We retake up the problem discussed in example-1 - 2 and make decision about the selection of an optimal action through Bayesion principle. 0 The table of prior probabilities and state of nature q_i can be displayed as given below,

State of nature θ	θ_1=100	θ_2= 150	θ_3 = 200
A priori probabilities	0.5	0.4	0.1
Also the pay off table will be as below :			
States of nature	Prepare — 100(α1)	ActsPrepare - 150(α2)	Prepare – 200 (α3)
θ_1 = 100	100	- 150	-400
θ_2 = 150	400	150	-250
θ_3 = 200	700	450	200

The expected payoffs for each act a_1, a_2, and as can be computed in the following manner.

B (a1)= -100 x 0.5 + 400 x 0.4 + 700 x 0.1 = 280

B (a2) = -150 x 0.5 + 150 x 0.4 + 450 x .1 =30

B (a3)= -400 x 0.5 -250 x 0.4 - 200 x .1 = -280

The EP for act a1 is maximum and hence the action ai i.e. prepare 100 items is the preferable. The reader can do the above exercise by preparing opportunity loss table and calculating EOL. I am sure, they will be compelled to choose the same action as through E.P

ii. Preposterior analysis: If the decision maker feels that his apriori probabilities are not fully reliable, he may try to obtain some more information about the states of nature. For this, he can collect information from a clairvoyant and make use of this information either to maximise his profits or minimise his losses. The expected profits by making use of clairvoyant's information is known as expected payoffs of perfect information (EPPI). EPPI is also often called expected value of payoffs under certainty. The perfect prediction by the clairvoyant reduces the opportunity losses due uncertainty to zero. The difference between EPPI -EP is called expected value of perfect information (EVPI); EVPI is the maximum amount which the decision make can to clairvoyant for perfect prediction. Here it is worth pointing out that EVPI=EOL under uncertainty for selecting an action.

In nut-shell, preposterior analysis leads are to decide whether it is profitable to acquire perfect prediction or not. Consider the maximum payoffs for states of nature *81 ,* ft* and*83* under the acts a1, a2, and a3. Then,

EPPI =100 x 0.3 + 400 x .5 + 700 x 0.2

= 30 + 200 + 140 = 370

Highest expected payoff under uncertainty is Rs.280. So,

EVPI = 370 - 280 = Rs.90

So the decision maker can at most pay Rs.90 to the forecaster.

Posterior analysis:

For posterior probabilities. Her borrow the example .8.7 of Basie Statistics authored by B.L. Agarwal.

Example : 1.7 : A professional economist, approaches the contractor to make use of his forest. The consultant actually does not tell the exact probabilities of a fixed percentage of a price rise, but only tells about the trend i.e whether there will be a fast rise in prices or whether prices will rise at the slow rate dun contract For brevity, we write the forecasts as fast and slow On The basis of this information, the economist also given the reliability statement for various price rise as, the probability of price rise in view of forecast 'fast' is 0.2 i.eP (Fast / 5%) = 0.2and for forecast 'slow' is 0.8 i.e.

P (Slow/ 5%) = 0.2

and for forecast ‘slow’ is i.e P (Slow/ 5%) = 0.8

For 8% price rise, the probability under forecast ‘fast’ is 0.7 i.e P (Fast / 8%) = 0.7 and 0.3 for forecast 'slow', P (slow / 8%) = 0.3. For10% price rise the probability under the forecast 'fast' is 0.9 i e. P(Fast / 10%) = 0.9 and 0.1 for forecast ‘slow’ i.e P (slow / 10%) = 0.1 we know that the prior probability assessed by the contractor in EVPI are

p(5 %) = P (5 % price rise) = 0.4

P(8 %) = P (8 % price rise) = 0.5

P(10 %) = P (10 % price rise) = 0.1

The posterior probabilities are given by,

$$P(\text{Event} \;/\; \text{forecast result})$$

$$\frac{P(\text{Event I forecast result})}{P(\text{Forecast result})}$$

By Bayes' formula (5.11) we obtain

$$P(\text{Event I Forecast}) = P\left(\frac{\text{Forecast}}{\text{Event}}\right) \times P(\text{Event})$$

We first calculate $P(\text{Event} \cap \text{forecast result})$ separated by formula (18.5) making use of the probabilities given statements and the p prior probabilities.

P (5% fast) = P(fast/5%) P (5%) = 0.2x0.4 = 0.08.

P (5 % slow) =P (slow/5 %) P (5 %) - 0.8 x 0.4 - 0.32

P (8% fast) =P (fast/8%) P (8 %) = 0.7x0.5 = 0.35

P (8 % slow) = P (slow/8 %) P (8 %) = 0.3 x 0.5 = 0.15

P (10 % fast) = P (fast/10 %) P (10%) = 0.9 x 0.1 - 0.09

P (10% slow) = P (slow/10 %)P (10%)= 0.1x0.1 = 0.01

The probability of price rise 'fast' = 0.08 + 0,35 + 0.09 = 0.52

The probability of price rise 'slow' = 0.32 + 0.15 + 0.01 = 0.48

The posterior probabilities by the formula (18,4.1) are,

$$P(5\%/\text{Fast}) = \frac{P(5\% \text{ I Fast})}{P(\text{Fast})} = \frac{0.08}{0.52} = 0.1538$$

$$P(5\%/\text{Slow}) = \frac{P(5\% \text{ I Slow})}{P(\text{Slow})} = \frac{0.32}{0.48} = 0.6667$$

$$P(8\%/\text{Fast}) = \frac{P(8\% \text{ I Fast})}{P(\text{Fast})} = \frac{0.35}{0.52} = 0.6731$$

$$P(8\%/\text{Slow}) = \frac{P(8\% \text{ I Slow})}{P(\text{Slow})} = \frac{0.15}{0.48} = 0.3125$$

$$P(10\%/\text{Fast}) = \frac{P(10\% \text{ I Fast})}{P(\text{Fast})} = \frac{0.09}{0.52} = 0.1731$$

$$P(10\%/\text{Slow}) = \frac{P(10\% \text{ I Slow})}{P(\text{Slow})} = \frac{0.01}{0.48} = 0.0208$$

It can be verified that the sum of the probabilities :

$$\sum P(\text{Event/fast}) = 0.1538 + 0.6731 + 0.1.731 = 1.0$$

$$\sum P(\text{Event/slow}) = 0.6667 + 0.3125 + 0.0208 = 1.0$$

Once we have got the posterior probabilities, there is no sense in confining ourselves to the use of prior probabilities. Hence, we calculate the expected monetary values under the forecast 'Fast' and 'Slow' superlatively, for a fixed price and cost plus percentage contract.

For the forecast 'Fast',

EMV (Fixed price)=22500 x 0.1538 + 6000 x 0.6731 – 5000 x 0.1731

= 3460.50 + 4038.60 – 865.50

= 6633.60

EMV (Cost plus) = 1550 x 0.1538 + 19880 x 0.6731 + 20100 x 0.7131

= 3006.79 + 13881.23 + 3479.31

= 19867.33

EMV for cost plus percentage contract is greater than EMV for fixed price contract in spite of the use of posterior probabilities under the new forecast fast. For the forecast 'Slow',

For the forecast 'Slow'

EMV (Fixed price) = 22500 x 0.6667 + 6000 x 0.3125 – 5000 x 0.0208

= 15000.75 + 1875.00-104

= 16771.75

EMV (Cost plus) = 19550 x 0.667 + 19880 x 0.3125 +20100 x 0.0208

= 13033.98 + 6212.50 + 418.08

= 19664.56

Again, for the price rise forecast 'slow', cost plus percentage contract is better than fixed price contract. We can also calculate the expected value, for selecting cost plus

percentage contract in both kind of forecasts, for contractor's problem the expected value is,

19867.33 x 0.52 + 19664.56 x 0.48 = 19770.00

To avoid confusion, it is worthwhile to point out, that a situation may arise in which a decision maker may select one course of action with the arise forecast 'Fast' and the other course of action with the forecast 'Slow'.

If the contractor has to pay fee to the consultant say Rs.500, this amount has to be deducted from the expected value. In the present example, there is no gain after paying any fee, and the contractor would not like to buy the forecast.

QUESTIONS

1. Payoff table showing profits (Lakhs of rupees) for various sizes of plants and demand levels is given below:

Demand units	Probability	20000 units	Profits (Lakhs Rs.) Plant capacity 30,000 units	40,000 units
10,000	.3	-4.0	- 6.0	- 8.0
20,000	.4	1.0	0	- 2.0
30,000	.2	1.5	5.0	4.0
40,000	.1	2.0	6.5	7.5

What capacity plant should he install?

2. Keeping in view the demand of a product, a manufacturer wants to explore whether he should continue with present plants or expand it or replace the old plant with the new one. The market position with expected payoffs is as tabulated below.

Sales	Probability of Sales	Maintain old plan	Expend the plant	Install new plant
< 20,000	.2	30	-30	- 40
20,000 to 40,000	.4	40	-20	0
> 40,000	.3	50	50	60

Calculate the expected value of perfect information.

3. A shopkeeper costs Rs.3.00 per icecream and sells it for Rs.4.00 each. The demand pattern of icecreams per day with their respective probabilities is as follows:

No. of icecreams demanded	:	50	60	70	80	100
Probability	:	1	.2	.3	.15	.25

How many icecreams per day should the shopkeeper stock?

4. A businessman wants to construct a hotel. He is in a dilemma whether he should construct 50 rooms, 100 rooms or 200 rooms hotel. From a market study he could collect inform as follows:

		Monetary returns	
Size of Hotel	Good demand	Medium demand	Poor demand
50 rooms	50,000	30,000	10,000
100 rooms	80,000	60,000	20,000
200 rooms	1,00,000	90,000	40,000

(i) Give maximin decision

(ii) Give minimax decision

5. shipbuilding company has launched a programme for the construction of a new class of ships, certain spare units like the prime mover, each costing Rs.2,00,000 have to be purchased. If these units are not available when needed, a very serious loss is incurred which is of the order of Rs. 10,000,000 in each instance. Requirements of the spares with corresponding probabilities are given below:

No. of Spares	:	0	1	2	3	4	5
Probability of requirement	:	0.876	0.062	0.041	0.015	0.005	0.001

How many spares should the company buy in order to optimize inventory decision?

6. Two companies, Hindustan Electro-carbon Ltd., and Poly Chemicals Ltd., expect to announce plan for next year's operations on the same day. On vital issue that the shareholders of each company as well as the general public have an interest in, is the opposition that each of the companies will take regarding the problem of pollution. If one company, for example, declared its intent to take action towards stopping pollution, its public image will be greatly improved. But on the other hand such action could increase its costs and put it in a bad position with respect to its competitor, if the competitor chooses not to take the same course of action. Each company can take any of the following three actions:

1. Adoption of policy towards ending pollution

2. Complete avoidance of the issue, or

3. Intention to continue as in the past. The payoffs for the actions are:

Hindustan Electro - Carbon Ltd.	**Poly Chemicals Ltd.**		
	Action (i)	**Action (ii)**	**Action(iii)**
Action (i)	3	-2	4
Action (i)	-1	4	2
Action (iii)	2	2	6

Determine the optimal cause of action for each company.

7. The research department of Hindustan Lever has recommended to the marketing department to launch a shampoo of three different types. The marketing manager has to decide one of the types of shampoo to be launched under the following estimated payoffs for various levels of states.

Types of shampoo	Estimated levels of sales (units)		
	15,00	10,000	5,000
Egg shampoo	80	10	5
Clinic Shampoo	40	10	20
Delux Shampoo	55	15	3

What will be the marketing manager's decision if:

(i) Maximin (ii) Minimax (iii) Maximax (iv) Laplace (v) regret criterion is applied.

8. The estimated sales of proposed types of perfumes are as under:

Types of perfumes	Estimated levels of sales		(Units)
	Rs. 20,000	Rs. 10,000	Rs. 2,000
A	25	15	10
B	40	20	5
C	60	25	8

Make decisions under Minimax and Laplace method.

9. YZ Co. Ltd. wants to go in for a public share issue of Rs.10 lakhs (1 lakh shares of Rs.10 each) as a part of effort to raise capital needed for its expansion programme. The company is optimistic that if the issue were made now it would be fully taken up at a price of Rs.30 per share.

However, the company if facing two crucial situations, both of which may influence the share prices *in* the near future, namely:

i. An impending wage dispute with assembly workers which could lead to a strike in the whole factory could have an adverse effect on the share price.
ii. The possibility of a substantial business in the export market, which would increase the share price. The four possible event and their expected effect on the company's share prices are envisaged as:

El : No strike and export business obtained - share price rises to Rs.34.

E2 : Strike and export business obtained - Share price stays at Rs.30.

E3 : No strike and export business lost - share price hovers around Rs.32.

E4 : Strike and export business lost - share price drop to Rs. 16.

And the management has identified three possible strategies that the Company could adopt; viz.,

S1 : Issue 1,00,000 shares now

S2 : Issue 1,00,000 shares only after the outcome of (a) and (b) are known

S3 : Issue 50,000 share now and 50,000 shares after the outcome of (a) and (b) are known.

You are required to:

1. Draw up a payoff table for the company and determine the minimax regret solution. What alternative criteria might he used.

2. Determine the optimum policy for the company using the criterion of maximising expected pay-off, given the estimate that the probability of a strike is 55 % and there is a 65 % chance of getting the export business, these probabilities being independent.

3. Determine the expected value of perfect information for the company.

10. A group of volunteers of a service organisation raises Money each year by selling gift articles outside the stadium after a football match between Team X and Y. They can *buy* any of the three different types of gift articles from a dealer. Their sales are mostly dependent on which team wins the match. A conditional pay-off table is as under:

	Type of gift articles		
	I	II	III
Team X wins	Rs. 1,000	900	600
Team Y wins	Rs. 400	500	800

i Construct the opportunity Loss Table and

ii Which type of gift article should the volunteers buy if the probability of Team X winning is 0.8.

REFERENCES

Agarwal, B.L. : *'Basic Statistics',* Wiley Eastern Ltd., New Delhi, 2nd ed., 1991.

Berger, J.O.: *'Statistical Decision Theory* Springer- Verlay, Berlin, 1980.

Byrkit, D.R. : *'Elementary Business Statistics',* D.Vari Nostrand Company New York.

Hoel, EG. and Jessen, R.J : *'Basic Statistics for Business' arid Economies',* John Wiley, New York, 1982

Lapin, L.L. : *'Quantitative Method for Business Decision',* Harcoust Braco, Jovanovicls, New York, 2nd ed., 1981.

Richard, L.E. and Lacava, J.J. : *'Business Statistics',* McGraw-Hill Book Company, New York, 1978,

- End Of Chapter -

LESSON - 18

CHI - SQUARE TEST IN CONTINGENCY TABLE

Preamble

Testing of hypothesis is an important tool of Statistics. There are many test used in Statistical analysis. But some are more frequently used than others. Chi-square test is one of the most frequently used test. The reason being that it is applicable in a large number of sciences like Biology, Agriculture, Psychology Education, Management, etc. Chi-square test makes use of the Chi-square distribution, that is why it is called chi-square test. The chi-square distribution is utilised to determine the critical values of the chi-square variate at various level of significance.

Like other tests chi-square test also entails null and alternative hypothesis, two types of error in test of hypothesis leading to level of significance and power of the test, degrees of freedom. The details of these are omitted here.

Chi-square test is applicable to test the hypothesis about the variance of a normal population, test of goodness of fit of a theoretical distribution, test of independence of attributes when the frequencies are presented in a two way table according to two attributes classified in various categories known as the contingency table.

Chi-square test dated back to 1900, when Karl Pearson gave the test statistics for frequencies classified into K-mutually exclusive categories.

Chi-square Statistic

Suppose there are K mutually classes and theoretically it is expected that they are likely to occur in the ratio

$$r_1 : r_2 : r_3 :... r_k$$

Let $O_1, O_2, O_3,.....O_k$ be the observed frequencies in k classes $C_1, C_2, C_3,....C_k$ respectively. Also suppose $E_1, E_2, E_3, ...E_k$ are the expected (theoretical or hypothetical) frequencies under null hypothesis calculated by the formula

$$Ei = \frac{ri}{r} x\ n \quad for\ i = 1,2.....,k \quad where \sum_{i=1}^{k} ri = r$$

$$and \sum_{i-1}^{k} Oi = n$$

Karl Pearson's chi-square statistic is,

$$\approx X^2 = \sum_{i=1}^{k} \frac{(Oi - ki)^2}{Ei} \qquad (2.1)$$

X^2 has(k-1) degrees of freedom(d.f)

Properties

i The value of Chi-square varies from 0 to ∞

ii When each $O_1 = E_i$, the value of Chi-square is zero

iii Chi-square can never be negative.

Contingency table

The data are often based on the count of persons, items, units or individuals which possess certain attributes. Here we categories individuals according to attributes say, A and B. Suppose the attribute A has on categories and P has q categories. For example, the attribute A represents Father's height categories and as tall, medium, gnome and attribute B as son's height categories as tall, medium and gnome. The frequencies are to be observed for all combination of Father's height and Son's height.

A contingency table with attributes A and B having p and q categories is displayed below. Attribute A is taken along rows and B along columns. O_{ij} is the frequency of $(ij)^1$ cell which represent the number of individual in a group which possess the attributes Ai and Bj for i = 1, 2,... p and j = 1,2,... q. The contingency table with p >2 and/or q >2 is called a manifold contingency table. A contingency table with 'p' rows and q columns is known as the contingency table of order (px q).

1 -1 Contingency table

Attribute A	Bi	Attribute B B2……Bj	B_q
A_1 A_2 . . . Ai . . . A_P	O11 O21 . . . Oil . . . O_{P1}	O12…….Oij O22……O2j O_{i2}……Oij O_P2……Opj	----------O ----------O . . . ----------O …….$O_{}$
Total	Cl	C_2……Cj	……

The contingency holds certain relations

$$\sum_{i=1}^{p} Ri = \sum_{j=1}^{q} cj = n$$

Where n is the total number of individual taken into consideration.

Also each of the row total or column total is known as marginal total.

Test of hypothesis in Contingency Table

Chi-square test is a test of independence of attributes in a contingency table. From the table 2-1, that a contingency is a rectangular array having rows and columns ascertained according to the categories of the attributes A and B.

The null hypothesis is,

Ho : Two attributes A and B are independent

Vs H1 : Two attributes are dependent on each other.

Test statistics for chi-square test in case of contingency table of order (pxq) is,

$$X^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(Oij - Eij)^2}{Eij} \qquad (2.2)$$

Statistic χ^2 has (p-1) (q-1) d.f.

The observed frequency O_{ij} is already available in the contingency table. Now the question remains to obtain the expected frequency E_{ij} corresponding to each O_{ij}.

Under H_0, the independence of attributes, the expected frequency,

$$Eij = \frac{ith\ row\ total \times j\ jth\ column\ total}{n}$$

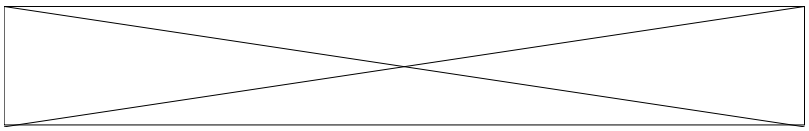
$$= \frac{Ri \times Ci}{n}$$

Once, all the expected frequencies are calculated, it is trivial to calculate the value of statistic chi-square by the formula (2.2). Here it should be checked that

$$\sum_i \sum_i Eij = \sum_i \sum_j Oij = n$$

Decision criteria:

The calculated value of chi-square is compared with the tabulated value of X^2 for (p-1) (q-1) d.f. and prefixed level of significance a . If the calculated value of X^2 is greater than the tabulated value of chi-square for (p-1) (q-1) d.f. and a level of significance, reject H_0. It means that the attributes are dependent on each other meaning that an unit possessing the attribute Bj. On the other hand if the calculated value of X^2 is less than the tabulated value of X for (p-1) (q-1) d.f. and a level of significance, then Ho is accepted. It shows the presence of attribute A has no bearing in the presence of the attribute B.



Please use headphones

Example 2-1:

Suppose a survey is conducted to know the opinion of the workers of a factory whether various types of incentives has got any relationship with category of worker or not. The data collected through the survey are displayed in the table below:

Incentive Schemes

Category of workers	Type I	Type II	Type III	Total
Labours	125 (85)	40 (34)	20 (66)	185
Scribes	160 (138)	65 (55.5)	75 (106.5)	300
Technical	65 (99)	50 (40)	100 (76)	215
Executive	110(138)	30 (55.5)	160 (106.5)	300
Total	460	185	355	1000

Ho : Choice of the type of incentive scheme is independent of the category of worker.

against

H_1: Choice of the type of incentive scheme is related to category of workers.

To test Ho we apply chi-square test. For this first we calculate the expected frequencies for all the cells by the formula (2.3). The expected frequencies computed below and are entered in the above contingency table in parenthesis. The frequencies are also rounded of to the nearest unit value.

$$E11 = \frac{185 \times 460}{1000} = 85.10 = 85$$

$$E12 = \frac{185 \times 185}{1000} = 34.22 = 34$$

$$E13 = \frac{185 \times 355}{1000} = 65.67 = 66$$

$$E21 = \frac{460 \times 300}{1000} = 138$$

$$E22 = \frac{185 \times 300}{1000} = 55.5$$

$$E23 = \frac{300 \times 355}{1000} = 106.5$$

Similarly,

E31 = 98.9 = 99; E32 = 39.78 = 40; E33 = 76.3 = 76

E41 = 138; E42 = 55.5 =56, E43 = 106.5 = 107

Here it should be kept in mind that the sum of expected frequencies for each row and column is the same as the, sum of the observed frequencies.

The value of statistic chi-square is,

$$
\begin{aligned}
X^2 &= \frac{(125-85)^2}{85}+\frac{(40-34)^2}{34}+\frac{(20-66)^2}{66}+\frac{(160-138)^2}{138}+\frac{(60-55.5)^2}{55.5}+\frac{(75-106.5)^2}{106.5} \\
&\quad +\frac{(65-99)^2}{99}+\frac{(50-40)^2}{40}+\frac{(100-76)^2}{76}+\frac{(110-138)^2}{138}+\frac{(30-55.5)^2}{55.5}+\frac{(160-106.5)2}{106.5} \\
&= 18.82+1.06+32.06+3.51+0.36+9.32+11.68+2.50+7.58+5.68+11.71+26.88 \\
&= 131.16
\end{aligned}
$$

Degree of freedom for the given contingency table of 4x3 is 3x2= 6. Let the prefixed level of significance a = 0.05. The reader's are referred to the appendix table VI of Basic Statistics authored by B.L. Agarwal. From the table of X^2 distribution, x^2 for 6 d.f. and 5% a level of significance is 12.59. The calculated value of $x^{2=}$ 131.16 is greater 12.59. Hence we reject H_0. Here we conclude that the choice of type of incentive scheme is associated with type of workers.

Example 2-2:

The following table gives the number of breakdowns of three machines in three shifts during a month.

		Machines		
30		A	B	C
40	Shift I	9(9)	10(9)	11(12)
30	Shift II	9(12)	12(12)	19(15)
100	Shift III	12(9)	9(9)	9(12)
	Total per machine	30	31	39

The hypothesis that the number of breakdowns on machines is independent of shift or not can be tested by chi-square test.

Ho : The breakdowns in machines are independent of shifts.

H1 : The breakdowns are subject to shifts.

The hypothesis Ho can be tested by chi-square test.

To calculate the value of statistic chi-square we work at the expected frequencies.

$$E11 = \frac{30\times 30}{100} = 9, \quad E12 = \frac{30\times 31}{100} 9.3 = 9, \quad E13 = \frac{30\times 39}{100} = 11.7 = 12$$

$$E21 = \frac{40\times 30}{100} = 12, \quad E22 = \frac{31\times 40}{100} = 12.4 = 12, \quad E23 = \frac{40\times 39}{100} = 15.2 = 15$$

Similarly,

$E_{31} = 9$, $E_{32} = 9$, $E_{33} = 12$

The expected frequencies are displayed in the contingency table itself in parentheses.

The value of statistic X" by the formula (2.2) is,

$$\begin{aligned} X^2 &= \frac{(9-9)^2}{9} + \frac{(10-9)^2}{9} + \frac{(11-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(19-15)^2}{15} + \frac{(12-9)^2}{9} \\ &\quad + \frac{(9-9)^2}{9} + \frac{(9-12)^2}{12} \\ &= 0 + 0.11 + 0.08 + 0.08 + 0 + 1.06 + 1 + 0 + 0.75 \\ &= 3.08 \end{aligned}$$

For the given contingency table of order (3x3), the degrees of freedom for the statistic X is 4. Tabulated value of chi- square for 4 d.f. and a = 0.05 from appendix table VI of Basic Statistics by B.L. Agarwal is 9.48 which is greater than the calculated value of X = 3.08. Hence, we accept H_0. It leads to the conclusion that the breakdowns have nothing to do with the shifts.

CONTINGENCY TABLE OF ORDER 2X2

There are many situations in which the contingency table has 2 rows and 2 columns. In case of contingency table of order 2x2, short cut method i.e., a direct formula for chi-square can be used. This formula is derived from direct approach.

Suppose the contingency table of order (2x2) is,

	B_1	B_2	Total
A_1	a	b	a+b
A_2	c	d	c+d
Total	a+c	b+d	a+b+c+d =n

Where n is the sample size. One can of course calculate the value of chi-square fry calculating the expected frequencies. But by algebraic manipulation, the direct formula for chi-square statistic is

$$X^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (2.4)$$

X^2 has 1 d.f

Decision about the independence of the attributes A and B can be taken in the usual way i.e. if c^2 cal $>c^2$ reject H_0. It means that the attributes A and B are dependent. Again if c^2 cal, 2 a 1accept H_0. It leads to the conclusion that the attributes A and B are independent.

Example 2-3 : A sampled group of 50 persons was vaccinated to prevent against malaria. Also a control group of 50 persons was observed in the same colony. Following results were obtained.

	Suffered from	Not suffered from	Total
Vaccinated	12	38	5050
Not	35	15	
Total	47	53	100

Can it be concluded that vaccination checks malaria.

Here we test,

Ho : Vaccination has nothing to do with the prevalence of malaria against Hi: Vaccination prevents the occurrence of malaria. The value of chi-square by the formula (2.4) is,

$$X^2 = \frac{100(12\times15-35\times38)^2}{50\times50\times47\times53} = \frac{100\times1150\times1150}{50\times50\times47\times53} = 21.23$$

Tabulated value ofc2 for 1 d.f. and a = 0.05 level of significance is 3.841 which is

less than the calculated value of c2 = 21.23. Hence, we reject H_0. It leads to the conclusion that vaccination prevents malaria.

Example 2-4:

In a survey of 200 children of which 80 were intelligent, 40 has skilled fathers, while 85 of the unintelligent children had unskilled fathers. Do these information support the hypothesis that skilled fathers have intelligent children.

Firstly, we tabulate the data in a contingency table of order 2 x 2 as given below

	Fathers		
Children	Skilled	Unskilled	Total
Intelligent	40	35	75
Unintelligen	40	85	125
Total	80	120	200

The hypothesis,

Ho : Intelligence of children is independent of skill of fathers

against H_1 : Skilled fathers have intelligent children can be tested by c2- test.

The value of statistic

$$X^2 = \frac{200(40\times85-40\times35)}{80\times120\times75\times125} = \frac{200\times40\times40\times50\times50}{80\times120\times75\times125} = 8.9$$

Calculated c2 = 8.9 is greater than the tabulated value of chi-square for 1 d.f. and a = 0.5 i.e. 3.841. Hence, we reject H_0. It means that skilled fathers have intelligent children.

Yate's correction

Chi-square is a continuous distribution. Hence, the continuity criterion should be obtained. It has been observed that in a contingency table of order 2 x 2 if any of the cell frequency is small, say less than 5, the continuity is disturbed. Hence, Yates suggested a correction for continuity. The correction is that add 0.5 to the small cell frequency and add and subtract 0.5 from other cell frequencies in such a manner that the marginal totals remains the same. Then calculate chi- square in the usual manner with the adjusted frequencies by the formula (2.4).

Alternative approach

Instead of adding and subtracting 0.5 from the cell frequencies, a formula for chi-square has been developed which has emerged after making the adjustments. In this way, the botheration of adding and subtracting 0.5 and dealing with the fractional frequencies is avoided. The formula for the contingency table of order (2x2).

		Bl	B2	Total
	Al	A	B	(a+b)
	A2	c	d	(c+d)
	Total	a +c	b+d	n

is

$$X^2 = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

The quantity |ad-bc| represents the absolute value of the difference (ad-bc). It means whether the difference is positive or negative it has to be taken as positive.

As usual c2 has 1 d.f.

The decision about Ho is taken in the usual manner.

Here it will be worth pointing out that the value of statistic % obtained by adjusting the cell frequencies and calculatingx2 by the formula (2.4) is always same as we get directly by the formula (2.5)

Example 2-5 :

The following data give the number of persons classified by in a sample of 200 addicts habit of drugs addiction and survival after ten years.

Drugs			
Survival	Addicted	Not Addicted	Total
Dead Alive	117 3	13 67	130 70
Total	120	80	200

To test whether the survival and addition are independent or not, we use chi-square test. Since, a cell frequency will use Yates correction. Now we solve the above example by adjusting frequencies and also directly by the formula.

To implement Yate's correction, we add 0.5 to the cell frequency 3 and subtract 0.5 from 117 and again add 0.5 to 13 and subtract from 67. In this way, the adjusted contingency table is,

Drugs			
	Addicted	Not	Total
Dead	116.5	13.5	130
Alive	3.5	66.5	70
Total	120	80	200

The test statistics c2 can be calculated by the formula (2.4) from the above table.

$$X^2 = \frac{200(116.5 \times 66.5 - 3.5 \times 13.5)^2}{130 \times 70 \times 120 \times 80}$$
$$= \frac{200(7747.25 - 47.25)^2}{130 \times 70 \times 120 \times 80}$$
$$= \frac{200 \times 7700 \times 7700}{130 \times 70 \times 120 \times 80}$$
$$= 135.7$$

Again the value of chi-square by the formula (2.5) from the given contingency table

is,

$$X^2 = \frac{200(1\ 117 \times 67 - 13 \times 31 - 100)^2}{130 \times 70 \times 120 \times 80}$$
$$= \frac{200 \times 7700 \times 7700}{130 \times 70 \times 120 \times 80}$$
$$= 135.7$$

The two approaches lead to the same value of statistic c2 Tabulated value of chi-square for 1 d.f. and 5 per cent level of significance is 3.841. The calculated value of x is greater than the tabulated value. Hence we reject Ho. It leads to the conclusion that addiction and survival are related to each other.

Example 2-6 :

The number of companies classified by licensed and unlicensed companies and the proportion of projects were as tabulated below:

Types of companies			
Proportion of projects	Licensed	Un licensed	Total
Up to 10 percent	2	6	8
More than 10	8	9	17
Total	10	15	25

The hypothesis

H_0 : Proportion of profit and licensor type are independent.

against

H_1 : Proportion of profit and licensor type are dependent.

can be tested by chi-square test.

Since a cell frequency is only 2, one have to make use of Yates correction for continuity.

Again we will calculate the statistic % by adjusting the frequencies as well as by direct formula first to show that both the approaches lead to the same value. The contingency table after adjustment comes out to be

	Lic.	Un Lic.	Total	
Up to 10 per cent	2.5	5.5	8	
More than	7.5	9.5	17	
Total	10	15	25	

The value of Statistic

$$X^2 = \frac{25(23.75-41.25)^2}{8\times17\times10\times15}$$
$$= \frac{25\times17.5\times17.5}{8\times17\times10\times15}$$
$$= 0.375$$

Now the value of chi-square by the direct formula is,

$$X^2 = \frac{25\left(12\times9-6\times8-\frac{25}{2}\right)^2}{8\times17\times10\times15}$$
$$= \frac{25\times17.5\times17.5}{8\times17\times10\times15}$$
$$= 0.375$$

c2cal = 0.375 is less than the tabulated value $c2_{.0591}$ = 3.841. So we accept Ho. We conclude that type of the company has no bearing on the proportion of profits.

Coefficient of contingency

Rejection of the independence of two factors reveals that the factors are associated with each other. But it fails to delineate the strength of dependence. This can be very well measured by coefficient of contingency The formula for coefficient of contingency is,

$$C = \sqrt{\frac{X^2}{X^2 + v}}$$

Where c2 is the calculated value of statistic chi-square and n is the sample size.

If the value of chi-square is zero, c = 0, If c2 is large and n is small, the value of c is near zero but never attains 1. If value of c is near zero, it shows a poor degree of dependence.

Again a value of chi-square nearing unity shows a high degree of dependency between the two factors. For a contingency table of order 5x5, the maximum value ofc is 0.894.

It should be kept in mind that if c2 test shows independence, coëfficiënt of contingency should not be calculated.

Exarnple 2-7:

We calculate the value of coëfficiënt of contingency for the example 2-1. The value of c2 = 131.16 and n = 12. So the value of

$$C = \sqrt{\frac{131.16}{131.16 + 12}}$$
$$= 0.916$$

coëfficiënt of contingency

The value of C = 0.916 is near to unity. Hence, there is a strong associaticn between the type of workers and incentive schemes.

QUESTIONS

1 A survey was conducted to investigate the views about the size of family from male and female groups separately Their opinion is tabulated below:

Sex	Small family	Medium size	Large family
Male	15	24	5
Female	7	22	7

Test the hypothesis that opinion about the size of family and sex are independent. If H_0 is rejected, fmd the value of coëfficiënt of contingency.

2. Sample of respondents classified by socia] class and political thinking is tabulated below:

Political Thinking	Poor	Middle	Elite

Communist	68	28	4
Socialist	24	106	20
Capitalist	10	34	56

Test whether the political thinking is associated with social class.

3 The candidates from rural and urban areas appeared in a competition and the results pertaining to their selection or non-selection were as follows:

Result			
Sex	Selected	Not selected	
Rural	40	110	
Urban	80	20	

Can it be inferred that the selection of candidates is associated with sex.

4 The following table provides the results of a survey on a group of persons to flnd out whether the smoking causes lungs cancer

Persons

Lungs Cancer	Smokers	Non smokers
Yes	30	8
No	3	9

Can it be concluded that smoking and lung cancer are independent.

5 Let 'A' represent new therapy and 'a' represent old one and let 'B' represent those who die and 'b' represent those who remain alive. The information about 500 subjects is put in the following table.

	A	a
B	28	72
b	112	288

Can it be considered the type ot therapy has Deanng on me survival 01 me persons.

6 The table given below shows the data obtair.ed during an epidemie of cholera:

	Attacked	Not attacked
Incoulated	42	232
Not-inoculated	106	748

Test the significance of inoculation in preventing the attack of cholera.

[Given X2.05 = 3.841 for 1 d.f, 5.991 for 2 d.f, 7.815 for 3 d.f.]

7. The following table reveals the condition of the house and the condition of the children.

Condition of children	Condition of house clear	Not clear	Total
Very clear	76	43	119
clear	38	17	55
Dirty	25	47	72
Total	139	107	247

Using the chi-square test, fmd out whether the condition of house effects the condition of children.

8. The following table gives the joint distribution of 120 fathers and sons with respect to hair colour.

		Father's Hair colour			
		Black	Grey	Brown	Total
Son's	Black	8	12	10	30
Hair	Grey	7	18	20	45
Colour	Brown	10	10	25	45
	Total	25	40	55	120

On the hypothesis of chance, find out if there is significant associatión between fathers and sons with respect to hair colour. Use coëfficiënt of contingency C.

REFERENCES

Agarwal, B.L.: *'Basic Statistics',* Wiley Eastern Ltd., New Delhi, 2nd ed., 1991.

Meyer, RL.: *Introductory Probability and Statis-tical Applications*,* Addison Wesley Publishing Company, Philippines.

Rehman, N.A.: *'Practical Exercises in Probability* and Statistics', *Charles Griffin,* London, 1972.

Walker, H.M. and : *'Statistical Inference',* Henry Hoft Lev, J. Company, New York, 1953.

- End Of Chapter -

LESSON - 19

MEASURES OF ASSOCIATION

Introduction

There are two types of studies, one regarding the variables and the other regarding attributes. The associatión between variables is usually made through correlation studies. Whereas the associatión between two attributes is measured through chi-square. Also the associatión between two or more attributes (qualitative factors) is measured through specially defined coefficients of associatión like Yule's coefficient, coefficient of colligation etc.

Here the point to emphasize is that correlation can be work out for the variables or characters which can be quantitatively measured. On the other hand associatión of attributes is applicable for those factors or characters where we can determine only the presence or absence of an attribute.

Notations

Since, we are dealing with the characters of units showing the absence or presence of attributes, various classes are to be formed. When we are dealing with one attribute say A, we can have one class denoted by A, showing the presence of attribute and the other by ‘a’ or’a’ , showing the absence of the attribute. If there are two attributes say A and B, there will be four classes formed out of the presence of the attributes A and B and their absence and *a* & *b.* The classes will be AB, A *a B &a b.* Similarly, three attributes A, B and C will have eight classes out of the presence of attributes A, B and C and their absence a, ,Y, namely, ABC, A BY, AbC, *a B C ,a b C* and *a b y* , . The frequencies of various classes are denoted by closing the classes in parentheses. For instance, the number of units showing the presence of the attribute A by (A) and those showing the absence of A by (a). Similarly frequency of the class Acr is denoted as (A/3) and of the class *a* as (a *(3)* and so on.

Terminology

Class frequency:

The number of units or individuals belongings to any class is known as class frequency. As stated the class frequency (B) denotes the number of units or subjects possessing the attribute B. The class frequency (Ab) denotes the frequency of class Ab i.e. the number of the units or subjects possessing the attribute A and not

possessing the attribute B. In the same manner any other class frequency can be defined.

Ultimate class frequency:

In the classification of units or subjects in a group under consideration, the frequencies of the classes of highest order are called ultimate class frequencies. For example, if there is one attribute considered for the units or subjects say A, then (A) and (a) are the ultimate class frequencies. For two attributes A and B, (AB) (Ab), (a B) and *(a b)* are the ultimate class frequencies. Similarly for three attribute A, B and C, the frequencies (ABC), (ABY), *(a* b Y).....are the ultimate class frequencies and so on.

Another point to emphasis is that any lower order frequencies can be expressed in terms of ultimate frequencies. For example, in case of two attributes, thé first order frequencies can be expressed in term of second order (Ultimate) frequencies i.e.

(A)= (AB) + (Ab)

(a)=(aB)+(ab)

(B)= (AB) + (aB)

(b) = (Ab) + (ab)

Similarly, for three attributes A, B, C

(A) = (ABC) + (AbC) + (ABy) + (Aby)

(a) = (aBC) + (ab) + (aBY) + (aj8y)

(B) = (ABC) + (ABy) + (aBC) + (aBy)

(b) = (AbC)+(Abg)+(abC)+(abg)

(C) = (ABC) + (A£C) + (aBC) + (a/3C)

(y) = (ABy) + (ABy) + (aBy) + (aby)

(AB) = (ABC) + (ABy)

(Ab) = (abC) + (abg)

(AC) = (ABC) + (AbC)

and so on.

Example 3-1:

Give the ultimate class frequencies for three attributes A, B and C as follows:

(ABC) = 120, (ABy) = 638, (AbC) = 220, (Aby) = 760, (aBC) = 200, (aBy) = 1162, (abC) = 161, (aby) = 1500

Calculate the frequencies of the classes A, B, C, AB, AC and BC.

We know, (A) = (ABC) + (AbC) + (ABy) + (Aby)

= 120 + 220 + 638 + 760 = 1738

Similarly,

(B) = (ABC) + (ABy) + (aBC) + (aBy)

= 120 + 638 + 200 + 1162 =2120

(C) = (ABC) + (AbC) + (aBC) + (abC) = 120 + 220 + 200 + 161 = 701

(AB) = (ABC) + (ABy)

= 120 + 638 = 758

(AC) = (ABC) + (AbC)

120 + 220 = 340

(BC) = (ABC) +(aBC)

= 120 + 200 = 320

Order of the class:

The order of a class depends on the number of the attributes involved in defining a class. A class having one attribute say A is known as the class of the first order. A class involving two attributes is called the class second order like Ab, aB, AB, Aa. Similarly, the classes like ABC, ABY, AbC etc. are called the classes of third order and so on.

Number of frequencies:

In a study of k attributes, the number of class frequencies is equal to 3^k . For instance,

for one attribute, the number of frequencies = 3^1 =3 since

k = 1 They are (A), (a), N.

For two attributes, the number of frequencies = 3^2 = 9. The nine frequencies for two attributes A and B of the positive, negative and ultimate classes can usefully be displayed in a two way table as follows:

	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	(αB)	β
Total	(A)	(α)	(N)

From the above table it is obvious that

A. = (AB)'+(Ab)

(a)= (aB) + (ab)

B. = (AB) + (aB)

(b)= (Ab) + (ab)

Note :

From the above we can make a general statement that any higher order frequency will never be greater than its lower order class frequency.

In like manner, we can give the class frequencies for three attributes A, B, C. The number of frequencies in this case will be $3^3 = 27$.

Inconsistency of data :

As a rule, no class frequency can be negative under any circumstances. If it happens, then the data are said to be inconsistent. For example, if any of the following inequality holds, the data will become inconsistent.

i) (AB) <0 implies (AB) will be - ve.

ii) (AB) > A implies (Ab) will be - ve.

iii) (AB) >B implies (aB) will be-ve.

iv) (AB) <(A) + (B) - N implies (ab) will be - ve.

When three attributes are under consideration and if any of the following inequalities holds, the data will be inconsistent.

i) (ABC) < O => (ABC) will be - ve.

ii) (ABC) < (AB) + (AC) - (A)=> (Ab g) will be - ve.

iii) (ABC) < (AB) + (BC) - (B) => (aBY) will be - ve.

iv) (ABC) < (AC) + (BC) - (C) => (abC) will be - ve.

v) (ABC) > (AB) => (ABY) will be - ve.

vi) (ABC) > (AC) => (AbC) will be - ve.

vii) (ABC) > (BC)=> (aBC) will be - ve.

viii) (ABC) > (AB) + (AC) + (BC)-(A)-(B)-(C)+N=> (abY) will be -ve.

In the process of finding out the measure of association if data are inconsistent, they are not suitable for use. Hence, either data should be corrected, if possible or rejected.

Consistency of data :

Contrary to inconsistency of data, it can be said that if any of the ultimate class frequency associated with the same population is not negative, the data are said to be consistent. It means that the data are in conformity with each other and are suitable for further analysis. For the test of consistency of data, the signs of inequalities given for inconsistency of data should be reversed and the word negative (-ve.) be changed to positive (+ve). Hence it can be stated that for consistency of data, no class frequency should be negative.

Example 3-2;

Given the class frequencies as below = (A)60, (b) =175, (AB) - 160, and N = 250

We can test whether the data is re-consistent or not by making the following two may table.

	A	a	Total
B	(AB) 160	(aB) -85	(B) = 75
b	(Ab) 220	(ab) -45	(b)=175
Total	A. = 380	a. = -130	250

In the above the frequencies for the class and are negative. Hence , the given data are inconsistent.

	A	a	Total

B	(AB)	(aB)	
	70	20	(B) = 90
b	(Ab)	(ab)	(b)=210
	155	55	
Total	(A)= 225	(a)= 75	N = 300

Since no class frequency is negative, we conclude that the data are consistent.

Example 3-4:

The following information was supplied by a tabulator about three attributes A, B and C.

N = 900, (A) = 200, (B) = 80, (C) = 10, (AB) = 180, (AC) = 140,

(BC) = 20 and (ABC) = 300.

The consistency of the data supplied by the tabulator can be tested by using the inequality

(ABC) > (AB) + (AC) + (BC) - (A) - (B) - (C) + N

if it holds, the data are consistent and if not, the data are inconsistent.

(AB) + (AC) + (BC) - (A) - (B) - (C) + N

= 180 +140 + 20 - 200 - 80 - 10 + 900

= 910

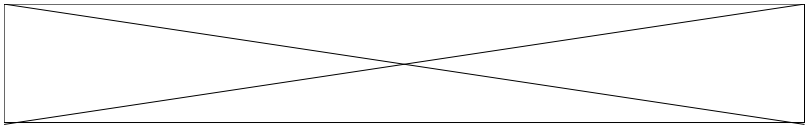
Given (ABC) = 300 is less than the value 910 of the class frequency (ABC) obtained from the inequality. Hence, the data are inconsistent.

Type of association

Two or more attributes may be positively or negatively associated or may be independent.

Positive association :

Two attributes A and B are said to be positively associated if the presence of an attribute is accompanied by the presence of the other. For instance, health and cleanliness are positively associated attributes.



Please use headphones

Negative association :

Two attributes are said to be negatively associated if the presence of one ensures the absence of the other.

Independence :

Two attributes are called independent if the presence or absence of one attribute has nothing to do with the presence or absence of the other.

If two events A and B are independent, then it is expected that the proportion of B's in A's is same as the proportion in a's and vice-versa i.e.

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)} \quad \ldots\ldots(3-1)$$

$$1 - \frac{(AB)}{(A)} = 1 - \frac{(\alpha \mathbf{B})}{(\alpha)}$$

$$(A) - (AB)/(A) = (\alpha) - (\alpha B)/(\alpha)$$

$$\frac{(A\beta)}{(A)} = \frac{(\alpha \mathbf{B})}{(\alpha)}$$

Similarly the proportion A's in B's same as inβ's. In this case, the relations that hold are:

$$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)} \quad \ldots\ldots(3.3)$$

and $$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \quad \ldots\ldots(3.4)$$

$$\frac{(AB)}{(\alpha B)} = \frac{(A\beta)}{(\alpha\beta)} \quad \ldots\ldots(3.5)$$

$$\text{or} \quad \frac{(AB)}{(B)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N} \quad \ldots\ldots (3.6)$$

$$\text{or} \quad (AB) = \frac{(A)(B)}{N} \quad \ldots\ldots (3.7)$$

$$\text{or} \quad \frac{(AB)}{N} = \frac{(A)}{N}\frac{(B)}{N} \quad \ldots\ldots (3.8)$$

$$(AB) = \frac{(A)(B)}{N}$$

Hence, if A and B are independent, if the proportion of frequencies of AB is equal to the product of proportion of frequencies in A and B.

Methods for measures of association

Various methods of association are as follows:

i. Frequency method
ii. Proportion method
iii. By Yule's coëfficiënt
iv. By Yule's coëfficiënt of colligation

(i) Frequency method

In this method we compare the observed frequency of the joint classes for two attributes A and E with their expected frequencies of the classes. If the observed frequency is greater than the expected frequency, then the association between the attributes is taken to be positive and if less, then negative. In case the observed frequency of the joint class is equal to its expected frequency, the two attributes are independent.

For two attributes A and B of a population of size N and their respective class frequencies (A), (B) and (AB). The expected frequency for the joint class (AB) by the-multiplicative law of probability under independence is

$$\frac{(A)}{N}\times\frac{(B)}{N}\times N=\frac{(A)(B)}{N}\quad(3.9)$$

So if $(AB)>\dfrac{(A)(B)}{N}$ +ve association

and if $(AB)<\dfrac{(A)(B)}{N}$ −ve association

Similarly for the joint class $\alpha\,\beta$,

if $(\alpha\ \beta)>\dfrac{(\alpha)(\beta)}{N}$ +ve association

and if $(\alpha\ \beta),\ <\dfrac{(\alpha)(\beta)}{N},$ −ve association

and so on.

The main drawback of this method is that we cannot find the degree of association between the two attributes. What we get is the kind of association only.

Example 3-5: -

Given the following data, find out whether the attributes A *nd B are independent positively associated or negatively associated.

Example 3-5: -Given the foUowing data, find out whether the attributes A *nd B are independent positively associated or negatively associated.

$$(A)=15,\quad (B)=25,\quad (AB)=30\quad\text{and}\quad N=50$$

$$\frac{(A)(B)}{N}=\frac{15\times 25}{50}=7.5$$

$$\therefore (AB)=30>7.5$$

Hence there is a positive association between A and B.

Example3-6: Given (A) = 130, (α) =240 , (AB) = 430, (βB) Find the kind of association between A and B.

We know

(A) + (α) = N = 310 + 240 = 550

(AB) + (αB) + B 430 + 180 = 610

Therefore (A) (B) = (310) (610)

N 350

= 343.8

(AB) = 430 > 343.8

It means A and B are positively associated.

Example 3-7:

Given the class frequencies as (AB) = 15, (Ab) = 25, (a B) = 45, (ab) = 5 find the type of association between A and B.

We know,

N = (AB) + (Ab) + (aB) + (aB)

= 15+25+45+5+ = 90

A = (AB) + (Ab) + 15 + 25 = 40

B = (AB) + (aB) + 15 + 45 = 60

$$\therefore \frac{(A)(B)}{N} = \frac{40 \times 60}{90} = 26.7$$

(AB) = 15< 26.7

Hence the attributes A and B are negatively associated.

(ii) Proportion method :

Two attributes A and B are said to be unassociated if the proportion of A's in B is same as amongst b's. Also if the proportion in/3 's then there is a positive association between A and B and if less than A and B are negatively associated. Symbolically, two attributes are related according as,

$$\text{If } \frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \text{ A and B are independent}$$

$$\text{If } \frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)}, \text{ A and B are positively associated}$$

$$\text{If } \frac{(AB)}{(B)} < \frac{(A\beta)}{(\beta)}, \text{ A and B negatively associated}$$

The above situations also hold good in the following situations also.

Independence : $\frac{(\alpha B)}{(B)} = \frac{(\alpha\ \beta)}{(\beta)}; \frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}; \frac{(AB)}{(A)} = \frac{(\alpha\ \beta)}{(\alpha)}$

+ ve association : $\frac{(\alpha B)}{(B)} > \frac{(\alpha\ \beta)}{(\beta)}; \frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}; \frac{(AB)}{(A)} > \frac{(\alpha\ \beta)}{(\alpha)}$

– ve association : $\frac{(\alpha B)}{(B)} < \frac{(\alpha\ \beta)}{(\beta)}; \frac{(AB)}{(A)} < \frac{(\alpha B)}{(\alpha)}; \frac{(AB)}{(A)} < \frac{(\alpha\beta)}{(\alpha)}$

In proportion method too, one obtains the kind of association only but not the degree of association.

Example 3-8:

In a study of 200 persons. 120 are poor. Out of 100 persons who suffered from T.B. 50 are poor. Find the kind of association between poverty and T.B.

Suppose A represents the attribute poor, B represent the attribute T.B. from the given question.

N - 200, (A) = 120, (B) = 100, (AB) = 50

From two way table, we can find other frequencies.

	A	α	
B	50	50	100
β	70	30	100
	120	80	200

$$\frac{(AB)}{(B)} = \frac{50}{100} = 0.5$$

$$\frac{(A\beta)}{(\beta)} = \frac{70}{100} = 0.7$$

$$\frac{(AB)}{(B)} < \frac{(A\beta)}{(\beta)}$$

Hence, the result reveals that A and B are negatively associated.

Yule's coefficient of Association

Yule's coefficient is the most popular measure of association. The main advantage of Yules coefficient is that it not only tells about the kind of association but also the degree of association. For two attributes A and B the class frequencies can be displayed in a two way table as given below:

Given : Number of runs up and down distribution table values

	A	α	
B	(AB)	$(\alpha\beta)$	(B)
β	$(A\beta)$	$(\alpha\beta)$	(β)
	(A)	(α)	(N)

For the cell frequencies given above, the coefficient of association between A and-B is given by the formula,

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3-10)$$

The value of Q lies between —1 and + l. If Q =1, there is a prefect positive association. If Q = -1, there is a perfect negative association

If Q = 0, the two attributes A and B are independent.

A high positive value i.e. greater than 0.5, shows a high degree of positive association between A and B. A negative value of Q less than 0.5 shows a high degree of negative association between A and B similarly other values of Q can be interpreted.

Example 3-9 :

We prepare two way table for the frequencies with two attributes A and B and calculate Yule's coefficient of associations given that

N m 800, (fl) = 330, (A) = 420, (AB) = 110 with the help of the given data, the two way frequency table is,

	A	α	
B	(AB)	$(\alpha\beta)$	(B)
β	$(A\beta)$	$(\alpha\beta)$	(β)
	(A)	(α)	(N)

For the cell frequencies given above, the coefficient of association between A and-B is given by the formula,

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3-10)$$

Yule's coefficient of association

$$110 \times 30 - 360 \times 310$$

$$Q = \frac{110 \times 30 - 360 \times 310}{110 \times 30 + 360 \times 31}$$

$$= \frac{-108300}{114900}$$

$$= -0.94$$

This shows that there is a high degree of negative association between A and B.

Example 3-10 : The distribution of the 600 persons according to their habit of smoking and alcohol drinking was as follows

	Alcohol drinkers (A)	Alcohol Nondrinkers (a)
Smokers (B)	340	60
Non- smokers (β)	80	120

We can find the association between smoking and habit of alcohol drinking by Yule's coefficient

$$Q = \frac{340\times120-80\times60}{340\times120+80\times60}$$
$$= \frac{40800-4800}{40 \quad 800+4800}$$
$$= \frac{36000}{45600}$$
$$= 0.79$$

Q = 0.79 leads to the conculsion that habits of smoking and alcohol drinking are highly positively associated

Coefficient of colligation:

Yule's gave another coefficient of association named as coefficient of colligation. It is denoted by Y. coefficient of colligation

$$Y = 1-\sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}$$
$$= 1+\sqrt{\frac{(A\beta)+(\alpha B)}{(AB)+(\alpha\beta)}}$$

There exist a relation between Q and Y. If is trivial to show that

$$Q = \frac{2y}{1+Y^2} \qquad \text{..... (3.11)}$$

Because of the inconvenience of calculations coefficient of colligation is less popular

All the more both the coefficient lead to the same results.

Example 3-11: For the problem given in example 3-10 we calculated the coefficient of colligation

$$Y = 1 - \sqrt{\frac{80\times 60}{340\times 120}} \Big/ 1 + \sqrt{80X60/340X20}$$

$$= \frac{1 - +0.343}{1+0.434}$$

$$= \frac{0.657}{1.343}$$

$$= 0.49$$

$$Q = \frac{2\times 8.49}{1+(.49)^2}$$

$$= \frac{0.98}{1+024}$$

$$= \frac{0.98}{1.24}$$

$$= 0.79$$

Here the value of Q is same as obtained in example (3-10)

Partial association :

The association between two attributes A and B may sometimes be due to the presence of a third factor C Hence it looks germane to find out the association between a and B in the sub populations C and Y, Thus the associations between A and B in the sub-populations C and y are called the partial associations and are denoted as QABy . The formulae of partial associations are

$$QABC = \frac{(ABC)(\alpha\beta C) - (A\beta C)(\alpha\beta C)}{(ABC)(\alpha\beta C) + (A\beta C)(\alpha\beta C)}$$

$$\text{and } QAB.y = \frac{(ABy)(\alpha\beta\gamma) - (A\beta\gamma)(\alpha\beta\gamma)}{(ABy)(\alpha\beta\gamma) + (A\beta\gamma)(\alpha\beta\gamma)}$$

Spurious association :

The association between two attributes A and B due to, some other factor are not considered which is known as spurious association.

QUESTIONS

I. Given that, N = 800, (A) - 200, (a B) = 40 and (AB) = 30

Test the consistency of data

2. Given the following frequencies for the three attributes.

A, B and C as,

(ABC) = 50, (AB) - 70, (BC) = 20, (B) = 28

Test whether the data are inconsistent.

3. According to a survey, the following results were obtained

	Boys	Girls
No. of candidates who appeared at an examination	800	200
Married	150	50
Married and successful	70	20
Unmarried and successful	550	110

Find the association between marital status and the success in the examination both for boys and girls.

4. In a class test in which 135 candidates were examined for proficiency in English and Economics, it was discovered that 75 students failed in English, 80 failed in Economics, and 50 failed in both. Find if there is any association between failing in English and Economics and also state the magnitude of association.

5. Give the following frequencies of the positive classes as, (A) = 950, (B) = 1100, (C) = 590, (AB) = 450, (AC) = 250, (BC) = 200, (ABC) = 120 and N = 10,000.

Find the frequencies of the remaining classes and test the consistency of data.

6. Out of 1000 people consulted, 811 liked chocolates; 752 liked toffees and 418 liked sweets, 570 liked chocolates and sweets and 348 liked toffees and sweets; 297 liked all three. Is this information correct?

7. Calculate the coefficient of partial association between A and B for the subpopulations C and from the following frequencies.

(ABC) = 300, (ABg) = 250, (AbC) = 210, (Abg) = 180,

(aBC) = 205, (abg) = 100, (abC) = 160 and (a/3y) = 175.

8. State the condition for two attributes A and B to be independent. Given the following information.

(AB) = 152, (aB) = 456, (Ab) = 26, (ab) = 78,

9. Prepare a 2´2 table from the following information and calculated Yule's coefficient of colligation

(A) = 100, (B) = 150, (AB) = 60, N = 500,

REFERENCES

Agarwal, B.L., 'Basic Statistic', Wiley Eastern Ltd., New Delhi, 2nd ed., 1991.

Ansari, M.A., Gupta, O.E and Chaudhari, S.S., 'Applied Statistics', KedarNath Ram Nath and Co., Meerut, 1980.

Garg, N.L., 'Practical Problems in Statistics', Ramesh Book Depot, Jaipur, 1978.

Goodman, L.A. and Krushal, W.H., 'Measures of Association for Cross Classification', Springer-Verlag, Berlin, 1979.

Gupta, S.C. and Kapoor, VK., 'Fundamentals of Mathematical Statistics', Sultan Chand, New Delhi, 7th ed., 1980.

Sancheti, D.C. and Kapoor, VK., 'Statistics', Sultan Chand, New Delhi, 1978.

- End Of Chapter -

LESSON - 20

NON-PARAMETRIC TEST

Introduction

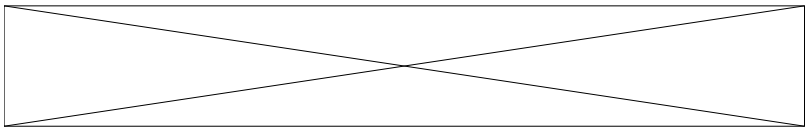
We are well versed with the theory of parametric tests as t, Z and F tests are most commonly used. But these tests are valid on certain assumptions. The most commonly used assumptions are that the variable follows normal distribution or the sample has come from a normal population. But the assumption of normality is not

always true. Even then people used parametric test freely under the umbrella of central limit theorem. If the distribution is highly skewed or the sample size is not large- enough to hold central limit theorem either the parametric test will not be reliable or the result obtained by them will be erroneous.

Usually the population parameters, mean and variance are estimated through sample values. But such an estimation will not be meaningful if the observation are on nominal scale like good, satisfactory, bad etc. or on ordinal scale. In this situation too, nonparametric tests yield good results.

The tests which do not require shape of the distribution are known as distribution free tests. The tests which do not depend on the parameters of the distribution like mean and variance are termed as nonparametric tests. In parlance or practice both these terms are used as synonyms.

Spearman's rank correlation was a breakthrough in the theory of nonparametric statistics which was well taken up by Harold hotel ling and later testing of hypothesis was initiated by Wilcoxon proposing a test for two sample cases. Anyhow, chi- square test is an interesting case which is categorized as parametric as well as nonparametric test. Now a large number of tests have been developed and nonparametric statistics is gaining ground day by day.



Please use headphones

Nonparametric statistics is largely used in Psychology, Education, Sociometry and other scientific phenomena where the observation are on nominal or ordinal scale. All the terms used in parametric tests like hypotheses, types of error, critical region, level of significance and degrees of freedom stand as such in their meaning and use in nonparametric test. Some people call the term level of significance as nominal in nonparametric tests.

Advantages of non-parametric tests

1. Nonparametric tests do not require any assumptions about the population distributions particularly about the normality of the population.
2. The calculations are simple.
3. Nonparametric tests are not complicated to understand.
4. Nonparametric tests are applicable to all types of data quantitative, nominal ordinal etc.
5. Many nonparametric tests are applicable even to the small samples.
6. Nonparametric tests are based on a few mild assumptions. Objectives

Generally, the hypotheses tested are about the median of the distribution of the population, the randomness of the population or whether populations have same or hypothetical distributions.

Problem of ties

If the variable is continuous there is no question of ties amongst the observations. But still due to limitations of measurements, rounding of figures etc., ties ac occur Two observations are said to be ties if they are equal. In ranking the observations, the problem arises how to award ranks to equal observations. The i^{th} group consisting of r_i tied observations

such k groups, then there will be $n \prod_{i=1}^{k} ri$! ways of orderings. The problem of ties can be surmounted in different ways, four of which are briefly given below:

(i) Midranks method:

In this method all the tied observations are arranged in order and ranked as if they are not tied. Then the average of the ranks of all the tied observations of a group is found out and each tied observation is given the same rank equal to the average value obtained. This is the most simple and frequency used method of dealing with the ties.

(ii) Average statistics method:

In this method the tied observations are arranged in all possible ways and ranked. The statistical value is calculated for each arrangement and then the average of these statistic value is used as the final value of the test statistic to take a decision about H0. But this method is not used since it needs two lengthy calculations.

(iii) Least favorable statistic method :

Instead of using the average value of statistics of all possible arrangements, one may choose a value out of all which minimizes the probability of rejection. This minimizes the probability of type I error.

(iv) Omitting the tied observations :

This is the most simple method but entails loss of information as the sample size is reduced by the number of tied values. The method can be used only when the number of tied values is small as compared to the sample size.

Assumptions about Non-Parametric tests:

Non-parametric tests are based on a few mild assumptions which are given below:

I. The first assumption is about the continuity of the distribution function. This is required to determine the sampling distribution.

2. Median is a good index of central tendency. In nonparametric test median is used as a measure of location parameter instead of mean. We know that mean. and median coincide in case of symmetric distribution.

One sample case:

Runs test:

It has been a common practice to write that the sample drawn is random. People standing in queue are in random order. Assumption of randomness may not come true in many cases. Hence, it becomes necessary to perform the test for randomness. Before we discuss runs test, it will be worth to discuss the runs first.

Definition of run:

A run is a sequence of like symbols preceded and followed by different kind of symbol(s) or no symbol(s).

Different runs may be exhibited by enclosing them in small vertical lines. A pattern of runs having a systematic arrangement of symbols shows the lack of randomness. For instance the observation in an order, all ladies ahead and all men behind them or one lady one gent in a queue show a lack of randomness. If we denote a lady by F and a gent by M, the sequences of the type, are considered to be nonrandom sequences. Whereas a sequence of the type,

M F M F M F M F

OR FFFFF MMMMMM

Are considered to non-random sequences. Whereas a sequence of the type

M M F F F M FFMMFMMMMFFF

is very likely a random sequence.

In the first sequence, the number of runs r = 8

In the second sequence, r = 2

In the third sequence, r = 8

As a general principle too many runs or a few runs are a mark of no randomness. Whereas an adequate number of runs in a sequence confirms randomness.

Now we come back to runs test.

Test for randomness :

Now we consider a problem of test of randomness in case of one sample of size n having two kinds of symbols a and b numbering ri and n2 respectively. The null

hypothesis, H0: the symbols a and b occur in random order in the sequence against the alternative, Hi: symbols a and b do not occur in random order, can be tested by the run test. Let the sample of size n contains n1 symbols of one type, say a, and n2 symbols of the other type, say b. Thus, n = n1 + n2. Also, suppose the number of runs of symbol are ri and that of symbol b are r2. Suppose ri + r2 == r. In order to perform a test of hypothesis based on the random variable R, we need to know the probability distribution of R under H0.

The probability distribution function of R is given as,

$$fR(r) = \binom{n1 - 1}{r/2 - 1}\binom{n2 - 1}{r/2 - 1} / \binom{n1 + n2}{n1}$$

When r is even

For r even the number of runs of both types must be the same i.e r1 = r2 = r/2 Again

$$fR(r) = \left[\binom{n1 - 1}{\frac{r}{2} - 1}\binom{n2 - 1}{\frac{r}{2} - 3} + \binom{n1 - 1}{\frac{r}{2} - 3}\binom{n2 - 1}{\frac{r}{2} - 1}\right] / \binom{n1 + n2}{n1}$$

When r is old

For r odd, r1 = r2 ± 1. In this situation, the sum is taken over two pairs of values, ri = (r - l)/2 and r2 = (r+l)/2 and vice versa.

Decision criteria :

To decide about Ho, the observed value of number of runs is compared with the lower and upper, critical number of runs. The critical values of the number of run can be seen in appendix Tables XI-Fi and XI-Fii of Basic Statistics by B.L. Agarwal at level of significant and n1 and n2 symbols,

If the observed value of r lies in between the critical values, the hypothesis of randomness is accepted otherwise rejected.

Example 4-1 : Given the following sequence of letters a and b

aa bbb a a b b a b b b b a a a b

we test the null hypothesis H0: the symbols V and 'b' occur in random order.

For the given sequence,

$n1 = 8$, $n_2 = 10$ and $n = 18$

The number of sequences, r = 8.

From the tables, the critical values of lower and upper critical number of sequences are, n = 5 and r2 = 15. The observed number of runs = 8 which lies between 5 and 15. Hence we conclude that the symbols ‘a’ and ‘b’ are in random order.

Further remark : In cases where the observations are taken are after the other and we take it for granted that they are taken randomly. To make sure, the runs test can be applied even for quantitative observations.

This type of data is dichotomized by taking the deviation from the median. Then supossingly the positive difference is symbolised by 'a’ and negative difference by 4b In this way we have a sequence of a's and b's. Runs test can be applied in the usual way

Example 4-2 :

Suppose a sample of 16 observations is as follows:

21.6, 15.6, 18.4, 30.8, 25.9, 17.6, 31.4, 33.8, 9.5, 11.6, 28,9, 40.2, 12.7, 7.4, 19.6, 54.3

The hypothesis of randomness of the set of observation can be tested in the following manner.

Arrange the observation in ascending order to find median.

7.4, 9.5, 11.6, 15.6, 17.6, 18.4, 19.6, 21.6, 25.9, 28.9, 30.8, 31.4, 33.8, 40.2, 51.8, 54.3

The medium of the set of observations is $= \frac{21.6 + 25.9}{2}$

$= \frac{47.5}{2}$

$= 23.75$

Now taking the deviations and denoting the + ve and -ve differences by the symbols a and b, the sequence of symbols comes out to be,

b b b a a b a a b b a b b b a

In the above sequence the number of runs r = 8.

n1 = 7, n2 = 9

The critical values at a = .05 and n1 = 7 and n2 = 9 from the appendix tables XI-Fi and XI-Fii of Basic Statistics by B.L. Agarwal are 4 and 14. The number of runs in the sample i.e. 8 lies between 4 and 14. Hence, it can be said that the sample is a random sample.

Two samples case

Wold-Wolfowitz run test :

This test is applied to test the identically of two populations. If X and Y are two variable measuring the same character then we want to test,

Ho : Fx (x) = Fy (x) for all x

Hi : Fx (x) ¹ Fy (x) for some x

Here we shall be testing H0 against Hi on the basis of runs obtained from two samples drawn from the two populations. Let two samples be drawn from population X and population Y denoted by the distribution function Fx(x) and Fy (x) of size m and n respectively Suppose the two independent samples are

..... X_1, X_2 -----Xm

And ..... Y_1, Y_2 — Yn

Procedure :

Combine both the samples and arrange the observations in order (ascending order), under the assumption of continuous variables, no ties should occur. In the combined samples track be maintained which observation belongs to which sample. For the observation belonging to a particular sample should either be underlined or some other identification mark may be used. Let the combined sequence of ordered statistics with m = 7 and n = 8 as follows:

Y Y X X X Y X X Y Y Y X Y X Y

In the above sequence, there are 4 runs of ‘X’s and 5 runs of ‘Y's. In this way we have in all 9 runs.

Now we define a random variable denoting the number of runs in the combined sequence of m ‘X's and n ‘Y's. As we know, too few runs tend to reject Ho as too many runs do. Wald- Wolfowitz runs test of size has a critical region given by the inequality.

R < r_a

Where r_a is chosen to be the largest integer such that P (R o.

In two samples case, we have the letters X and Y instead 'a' and 'b' in one sample case. So R is same it was found in the case of one sample given by (3.1). In this case m ≡ m and n ≡ n2. The only difference between one sample and two samples runs test is that in Wald-Wolfowitz runs test, we use only one sided test where as in one sample case we use two sided test.

Decision criteria :

To decide about H0, we compare the value R = r with critical value of r at a level of significance and n1 and n2 (m and n) sample sizes obtained from the tables prepared by Swed and Eisenhart. The same table is reproduced in Table XI-F(i) of Basic Statistics by B.L. Agarwal. If the observed number of runs r is less the critical value of r for a level of significance and ni, n2 (m,n) d.f. we reject Ho- It means that the two populations are not identical from the view of randomness. If r >ra , ni, n2 we accept H0.

Large sample case

In case of large samples, we can apply normal deviate test. In the situation when m and n, both are greater than 10, a normal approximation may be made assuming

that $\frac{m}{m+n}$ and $\frac{n}{m+n}$ remain constant as $(m+n) \to \infty$. The variable R will be distributed with

$$\text{mean } (R) = \frac{2mn}{m+n} + 1 \qquad (4.2)$$

$$\text{and var } (R) = \frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)^2} \qquad (4.3)$$

Thus the normal deviate,

$$Z = R - \text{mean } \frac{(R)}{\sqrt{\text{Var } (R)}} \qquad (4.4)$$

Where Z -N (0,1)

Decision criteria : If the calculated value of z is greater that Zα, which is 1.96 for 5% level of significance, reject Ho. It means two population are different. Again if Z < Zα, it can be concluded that the two populations are identical.

Decision criteria :

If the calculated value of z is greater that Za, which is 1.96 for 5% level of significance, reject Ho. It means two population are different. Again if Z < Za, it can be concluded that the two populations are identical.

Problem of ties:

Theoretically, no ties should occur. Anyhow if they occur within the sample, it causes no problem. But if they occur across samples, there arise a problem. The solution to the problem of ties in this case is to break the ties in all possible range ways and computer 'r' in each case. Choose the largest value of r to take a decision about H0. Largest value of 'r' is preferred due to the fact that, this is the one value which will be least favourable to reject H0.

Example 4-3 :

A sample each was drawn from two populations one of smokers and the other of nonsmokers. Their clinical tests were conducted to measure the effects on their blood, cells, lungs etc. and the scores were as follows:

Smokers : 45, 30, 35, 40, 27, 32, 26

Nonsmokers : 20, 22, 23, 25, 21, 42, 19, 18

whether there is a significant difference between the score of smokers and - nonsmokers population can be tested by Wald-wolfowitz tests in the following manner.

Here we test.

Ho : The populations of smokers and nonsmokers are identical.

H_1 : The scores of population try to cluster.

To test Ho against Hi, we combine the two samples and put them in an ordered sequence.

18 12 20 21 22 23 25 26 27 30

32 35 40 45 42

The total number of runs in the combined sequence is three. Compare it with the critical value of R at a = 0.05 and n1 = m = 7 and ni2 = n = 8. The critical value of r = 4. The calculated value of r is less than the critical value. Hence, we reject Ho. This shows that the scores of smokers and nonsmokers are not identically distributed.

Example 4-4 : The production of an ore during the months of the years 1990 and i.991 were as follows:

Production (Tonnes)

Month	1990	1991
Jan.	105.4	95.4
Feb.	110.8	114.6
Mar.	115.0	116.9
Apr.	95.8	103.2
May.	99.6	98.4
Jun.	103.4	86.5
Jul.	118.7	98.8
Aug.	72.8	82.5
Sep.	90.2	104
Oct.	112.1	117.2
Nov.	122.5	123.4
Dec.	126.2	115.8

Can it be regarded that the production of ore during months in two years is a random process.

The hypothesis,

Ho : Production of ore is a random process.

H_1 : Production is related to months.

This can be tested by Wald-wolfowitz runs test. Since m and ore both 12 which are more than 10, we will apply large sample test.

Now we combine the samples and arrange them in an ordered sequence (ascending order).

72.8 82.5 86.5 90.2 95.4 95.8 98.4 98.8 22L£

103.2 103.4 104.6 105.4 110.8 112.1 114.6

115.0 115.8 116.9 117.2 118.7 122.5 123.4 126.2

The number of runs, r = 17

Here m = 12, n = 12. Thus by the formulae (4.2) and (4.3),

$$\text{Mean } (R) = \frac{2\times12\times12}{12+12}+1$$

$$= \frac{2\times12\times12}{24}+1$$

$$= 13$$

$$\text{Var } (R) = \frac{2\times12\times12(2\times12\times12-12-12)}{(12+12)^2(12+12-1)^2}$$

$$= 5.7$$

$\therefore$ The statistic

$$Z = \frac{17-13}{\sqrt{5.7}}$$

$$= \frac{4}{\sqrt{2.387}}$$

$$= 1.67$$

The calculated value of z = 1.67 is less than the tabulated value of z = 1.96 at 5% level of significance. Hence, we accept that the production of ore during the months is a random process.

Runs up and runs down test

Before we describe the test, it looks logical to define first the runs up and runs down. In this method the magnitude of each observation of a set is not compared with a single value but with the one immediately preceding it in a given sequence of observations. If the preceding value is smaller, 'a' runs up started and if is greater a runs down commences. Let a runs up is denoted by + ve sign and ‘a’ runs down by - ve sign. So the sequence of +ve and -ve signs is reflected in the form of runs. For example, if we consider a sequence of seven observations as 5, 8, 3, 9, 6, 7, 10. The sequence of runs up and runs down will be + - + - + +. In the given sequence, there are fair runs up and two runs down.

Test procedure

The hypothesis

Ho : sequence is random

H_1 : sequence is not random

can be tested by runs up and runs down test. From the above discussion it is clear that if there are n observations there will be in all (n-1) runs up and runs down i.e.(n-l) +ve and -ve signs.

Let us suppose that there are V runs in the sequence of + ve and +ve signs.

The decision about Ho can be taken by comparing the value of the probability obtained from Table M of Non parametric methods for quantitative analysis by

J.D.Gibbons with predecided level of significance a . Usually we take a = 0.05 or 0.01. If the tabular probability for n and V is less than a , we reject Ho otherwise we accept H0.

Note : For two tailed test the probability obtained by table M should be doubled.

Example 4-5 :

For the observations

5, 8, 3, 9, 6, 7, 10

The hypothesis

Ho : sequence is random

Against H_1 : sequence is not random

Can be tested by the runs up and runs down test in the following manner.

The runs up and runs down are

1 + 1-1 + 1-1+1+1

Here n = 7, V = 5

The probability P for n = 7, V = 5, by Table M as referred above is. 4417. Thus, 2 P = .8834, which is greater than predecided level of significance a = 0.05. Hence, we accept HG. It means the sequence is random.

QUESTIONS

l. In a queue for a coming bus in Delhi, the following arrangement of adults according to sex was observed:

M F M M M M F F M M M F F F F M M

On the basis of the above sequence of males and females denoted by M and F respectively, can it be concluded that there was a lack of randomness.

2. The test scores gained by boys and girls in an oral examination were as follows:

Boys (X) : 26, 25, 28, 30, 35, 42, 18

Girls (Y) : 32, 38, 22, 27, 9

Test by Wald-wolfowitz test whether the distribution of scores earned by boys and girls is identical.

3. A row of cotton plants in a field had the following sequence of healthy (H) and diseased (D) plants.

H H D H H H D D H D D D H H H H

Test the hypothesis of randomness against the alternative of clustering (one tailed test).

4. The number of births occurring on seven days of a week in a hospital were as given below:

Births (X): 15 22 38 28 18 25 18

By the method of runs up and runs down, test that the number of births during the week days is a random process against clustering. (Use one tailed test).

Given: Number of runs up and down distribution table values

N	V	Left Tail (P)	V	Right Tail (P)
7	1	.0004	6	.1079
	2	.0250	5	.4417
	3	.1909	4	.8091

5. The test scores in shooting competition were as given below out of 500 shots.

210, 305, 265, 405, 365, 410, 385, 390

295, 230, 175, 160, 435, 475, 290, 110

Test that the sequence of scores is random.

6. The efficiency of eight workers in a factory was measured in terms of time consumed by them in completing a task before and after the lunch.

Time in (Minutes)

Before Lunch: *8.6,* 11.5, 12.2, 9.6, 10.4, 12.4, 13.5, *8.8*

After Lunch : 10.2, 13.5, 10.5, 12.5, 11.6, 15.2, 16.2, 17.5

Test the combined ordered sequence for randomness against non-randomness.

REFERENCES

Agarwal, B.L., *'Basic Statistics',* Wiley Eastern Ltd., New Delhi, 2nd ed., 1991.

Daniel, W.W., *'Applied Nonparametric Statistics',* Houghton Miffin Company, Boston, 1978.

Gibbon, J.D., *'Nonparametric Methods for Quantitative Analysis',* Holt, Rinehart and Winston, 1976.

'Nonparametric Statistical Inference', McGraw-Hill Kogakusha Ltd., Tokyo.

- End Of Chapter -

LESSON - 21

TIME SERIES AND DETERMINATION OF TREND

Introduction

Economic system is dynamic and it changes with time. The study of national income, national production, demand, supply, wages, etc. more with time. So the data pertaining to economic variable(s) collected in 3 chronological order is called a time series. The study of movement of economic factor(s) in a chronological order leads to draw many conclusions for present policy making, planning and management. Also the estimates can be made well in advance so as to regulate our production, supply and prices etc. A large number of Economists and Statistician defined time series in their own way. Firstly we give the definition for time series alone.

1. Kenny-keeping: A set of data depending on the time is called a time series.
2. D Ya-lun Chau : A time series may be defined as a collection of readings belonging to different time periods of some economic variable or composite of variables.
3. Ceril HJVÏeyers : A time series may be defined as a sequence of repeated measurements of a variable made periodically through time.
4. P.G.Moore : A series of values over a period of time is called a time series.
5. W.Z. HirSeh : A time series may be defined as a sequence of repeated measurements of a variable made periodically through time.

Analysis of time series

The study of a time series to evaluate the changes occurring over time and to look for the causes for these changes is known as the analysis of time series. Analysis of time series defined by various workers is also quoted below:

1. W.Z. Hirsch : A main objective in analysing time series is to understand, interpret and evaluate changes in economie phenomena in the hope of most correctly anticipating the cause of future events.
2. P.H. Karmel : The analysis of time series has developed in the main as a result of investigations into the nature and causes of those fluctuations in

economic activity called trade cycles. Economic theory has suggested various explanations of trade cycles. Analysis of time series has attempted to test the plausibility or otherwise of these theories. At the same time, such analysis may suggest new hypotheses for economic theorists to work on.

The analysis of a time series may be due to long term changes or short term changes. The factors responsible for short and long term changes should be studied separately. Also while analysing the time series data quantitatively, the analyst should always use his own judgement and logic before taking a decision. Now we study the components of a time series and their analysis.

Components of a time series :

An observation taken at a point of time consists of four components namely:

(i) Trend or secular trend (T)

(ii) Seasonal changes (S)

(iii) Cyclical changes (C)

(iv) Irregular movement (I).

The original data 'O' influenced by these four components can be expressed by the following model.

Multiplicative model:

The relation between O and T, S, C and I under this model is,

$O = TxSxCxI$ (1.1)

Addititive model

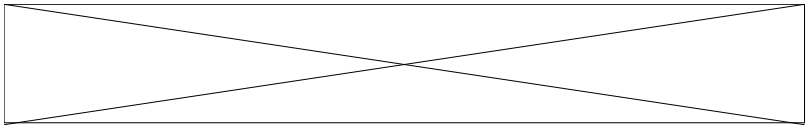
$O = T + S + C + I$ (1.2)

The analysis of time series is an attempt to segregate these component from the time series data and find their influence. If we remove the influence of any one or more, the data is left with the remaining one. If we eliminate the influence R, S and C, it is easy to infer that wherever variation is present in the data is due to irregular movement but not due to assignable causes like trend, seasonal or cyclical changes.

Secular trend :

Secular trend depicts the long-term movements, This is one of the main components of time series. It measures the slow changes occurring in a time series over a long period. For example, we find the rate of increase or decrease of food production over the last 20 years, we see the changes in GNP over the last two decades. Secular trend or simply trend does not take into consideration the short term fluctuations but only takes case of the long term fluctuations.

Trend is found out not only on national level but also by individual companies and industries to regulate their finances, marketing and production etc. There are various methods of estimating trend. They are named below:



Please use headphones

Graphical methods

i. Free hand method
ii. Semi-average method
iii. Moving average method

Mathematical method

i. Fitting of a straight line
ii. Fitting of logarithmic straight line
iii. Fitting of a Parabola

Editing of data :

Before we analyse a time series data, it is necessary to edit the data keeping in view the purpose of study. In the analysis, usually a value over time is compared with the other. So the data should be adjusted to ensure comparability. Generally, the data are adjusted for:

(i) calendar variations

(ii) price variations

(iii) population changes

(iv) miscellaneous changes.

We describe these changes in brief.

i. Calendar variation : Number of studies are related to consumption, production, demand during the months of a year. We know that the number of days vary from month to month. Hence, the monthly data have to be adjusted to the same number of days, say, 30 days, Such an adjustment is known as the editing of data for calendar variation.

ii. Price variation : The sales or quantity produced by a company cannot be judged correctly by considering only the monetary value. But it should be judged in terms of quantity given by the formula,

$$q = v/p \qquad (1.3)$$

where q = quantity of sales or production in a flxed period.

v = value of total product

p = price per unit during that period.

iii. Population changes : The consumption figures cannot totally be a criterion for demand of a particular item. But they have to be viewed in the make of population. The liking or demand for an item may be decreasing but the total demand may be increasing due to the increased population. Hence, the demand be measured in term of consumption per head. Hence, while comparing the fïgure or various periods in a time series adjustment for population changes be made in the data.

iv. Miscellaneous changes : While comparing the figures in a time series data, one should take care of type of items also. In the beginning there were no colour T.V sets. Hence to compare the production of coloured TV sets in value with black and white will not be proper. Hence, suitable adjustments should have to be made.

Also the units of measurements often change. As there are no miles or yards but the units are kilometer or meters or kilograms. Hence, while comparing the figures of two periods in which the units are not similar, the fïgure should be reduced to similar units.

Fitting of Trend

Freehand method:

This is one of the graphical methods of fitting a trend line. Here we plot the time series data on a graph paper by choosing suitable scales to represent time along X-axis and variate values on Y-axis. Once all the data are plotted in the form of dots or crosses on the graph, a straight line is drawn on the graph paper through a transparent scale in between the points in such a way that almost half of the points are above the line and half are below it and also as many points as possible lie on it. The line will be a good fit if the sum of all the vertical distances from the points to the line is zero. But this much of accuracy is not easily attainable. Usually the best line is fitted visually.

Freehand method is not very preferable as for the same date, the line of best fit will vary from person to person. In this way a trend line by free hand method pro vides rough estimates and is not suitable for predictions.

Example 1-1 :

Gross National Product (GNP) at 1980-81 prices for public administration, defence and other services is given below for the years 1977-78 to 1988-89.

Years	GNP
1977-78	2.7
1978 - 79	4.3
1979-80	7.3
1980-81	4.1
1981 - 82	3.5
1982 - 83	7.8
1983-84	3.6
1984 - 85	7.3
1985-86	7.5
1986-87	7.9
1987-88	8.0
1988-89	5.6

For the data given above, the trend line can be fitted on the graph by freehand method in the following manner.

Take years on abscissa at a distance of 1 cm and 1 cm = 1 unit of GNP on the ordinate. Then draw a trend line through a transparent scale as depicted in fig 21.1 below

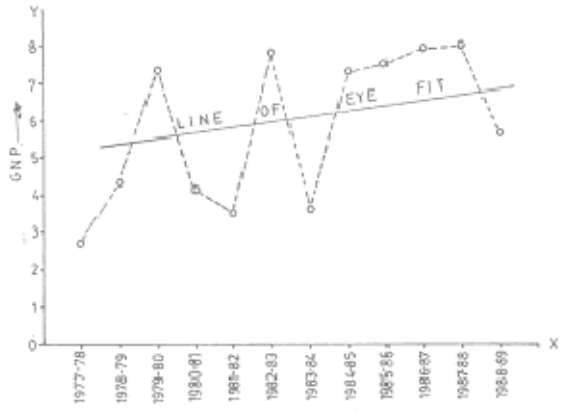


Fig.21-1 Trend by free hand method

Semi-average method :

The problems of drawing the trend merely by the judgement of the investigator is avoided in this method. The time series is divided into two equal halves consisting of the beginning half years and last half year and the average of each of the half series is calculated. The time series data are plotted in the usual way on a graph paper and also the average values are plotter against the mid periods of the corresponding half series. The points representing the average values are joined through a straight line and the line may be extended ". eyoiid points also. This line represents the trend line.

In the above process, one faces two types of problems.

(i) The number of time periods is odd and in this situation it is not possible to divide the series into two equal halves. The problem can be resolved by either including the middle most value in both the halves or neglecting it.

(ii) The number of years (periods) in half of the series is even. In that situation no year is mid-year of the half series. To resolve this problem, the average value should be plotted against the mid-point of the two middle years in both the half series.

Merit: Semi-average method has no subjectivity

Demerits : i) It is affected by extreme values if any

ii) It does not depict the true trend line.

iii) It does not ensure that the short term and long term fluctuations are eliminated.

Example 1-2 : For the data given in example 1-1, the trend line by semi-average method can be fitted in the following manner.

First half series consists of the years

$$= \frac{1}{6}(2.7 + 4.3 + 7.3 + 4.1 + 3.5 + 7.8)$$

$$= \frac{29.7}{6}$$

$$= 4.95$$

and that of last half series,

$$= \frac{1}{6}(3.6 + 7.3 + 7.5 + 7.9 + 8.0 + 5.6)$$

$$= \frac{39.9}{6}$$

$$= 6.65$$

Since the number of years in half series is six which is an even number hence no year will be a middle year of the series. The average values will have to be plotted against lst July 1979 and lst July 1985.

As per practice, the original data should be joined through dotted lines and the trend line by a smooth dark line.

The graph showing the trend line by semi-average method is given below.

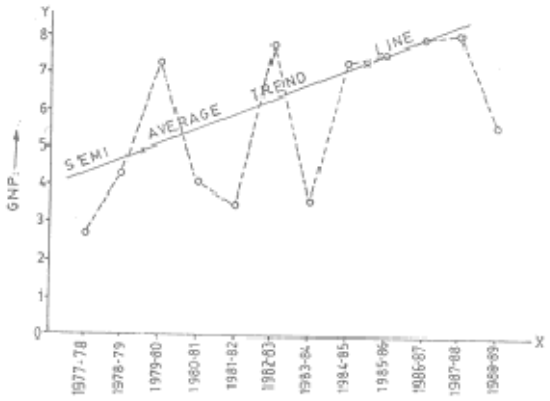


Fig.21.2 Trend by semi-average method

Moving average method :

The semi-average method cannot eliminate short term fluctuations except seasonals which are of little or no interest. But moving average method removes short term fluctuations as well. Moving average method is one of the most popular methods in time series analysis. Firstly we define a moving average.

Moving average is a series of average of the variant values corresponding to the sequences of fixed number of years (periods). The sequences are formed by deleting the first year of the last sequences and adding a subsequent year. The process continues till all the years are exhausted.

Now the question that arises is to determine the years to be taken to form the first group, so that the trend is reflected as a straight line. There is no hard and fast rule for it. As a principle, the minimum number of years be taken together in a group, which result in a straight line for the trend. As a thumb rule, the minimum number of years in a group, should be equal to the number of years (periods) which form a business cycle. An idea of cycles can be got by studying the short term fluctuations as adjudged by plotting the time series data. Once the number of years to be included in the first group is finalised, the method follows mechanically.

The moving averages are plotted on the graph paper by taking the averages along the Y-axis and years on the X-axis by choosing a suitable scale. The points so plotted are joined sequentially. The resulting gr provides the trend. If the points lie nearly on a straight line, the moving aw methods give a very clear picture of trend. On the contrary, if they do not fall in a line, we should search for a curvilinear trend. For a linear trend, the cycles should regular in amplitude and periodicity.

Merit:

The greatest advantage is the moving average method reduces the influence of extreme values.

Demerits :

i. The moving averages are not available for some of the beginning and end years. Hence, this method is not suit-able for projections.
ii. There is hardly any series which has regular cycles. Hence to use the same number of years for moving averages is not very logical.
iii. The method is not appropriate for the comparison of two series
iv. There is no hard and fast rule to decide the number of years to be taken in a group.
v. The method is not based on sound mathematical footing.

Example 1-3 :

For the data given in example 1.1, we will be fitting the trend by the method of moving averages taking 3 years in a cycle.

In the table below we reproduce the data along with the three year moving averages.

Years	G.N.P
1977 - 78	2.7
1978 - 79	4.3
1979 - 80	7.3
1980-81	4.1
1981 - 82	3.5
1982 - 83	7.8
1983-84	3.6
1984 - 85	7.3
1985-86	7.5
1986 - 87	7.9
1987 - 88	8.0
1988 - 89	5.6

$$\text{First moving average } \frac{2.7+4.3+7.3}{3}=\frac{14.3}{3}=4.77$$

$$\text{Second moving average } \frac{4.3+7.3+4.1}{3}=\frac{15.7}{3}=5.23 \text{ and so on}$$

The average value in the above table are plotted on the graph paper the years to which they are entered. The graph joining these point depicts the trend as shown in Fig.(21.3)

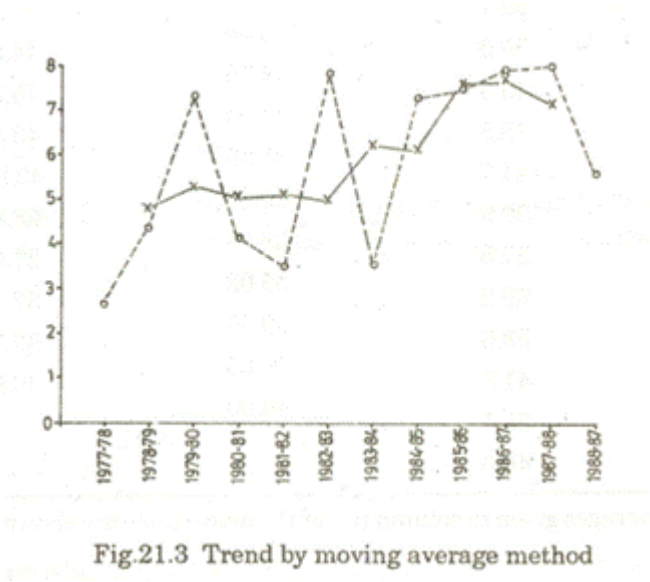


Fig.21.3 Trend by moving average method

The trend is not a linear.

Moving average method when the number of years in the cycle is even.

In this method the moving average is entered against the mid position of the two middle years. Since, in is moving average does not belong to any of the years given in the data. We again find out the moving average of the averages taking two averages at a time and enter them against the mid position which is a year of the given data. Once we have got the moving averages they are plotted on the graph paper in the usual way and trend is obtained by joining the plotted points sequentially.

Example 1-4 :

Net availability of per capita pulses per day (in gms) from 1972 to 1986 was as follows:

1972	1973	1974	1975	1976	1977	1978	1979
47.0	41.1	40.8	39.7	50.5	43.3	45.5	44.7
1980	1981	1982	1983	1984	1985	1986	
30.9	37.5	39.2	39.5	41.8	38.1	40.6	

Trend by moving average method taking four years moving averages is given below
Prepare the following table for moving averages.

Years (I)	Net availability of pulses per day (gms) (II)	4 - years moving average (III)	Mo
1972	47.0		
1973	41.1	42.15	
1974	40.8	43.02	
1975	39.7	43.58	
1976	50.5	44.75	
1977	43.3	46.00	
1978	45.5	41.10	
1979	44.7	39.65	
1980	30.9	38.08	
1981	37.5	36.02	
1982	39.2	39.50	
1983	39.5	39.65	
1984	41.8	40.00	
1985	38.1		
1986	40.6		

The moving average given in column (iv) of the above table are shown in the figure 21.4

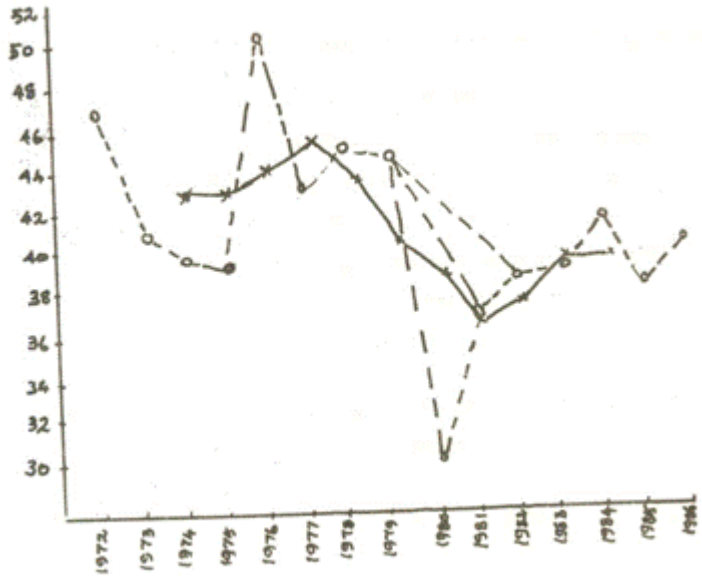


Fig.21.4 Trend by 4 years moving averages

A curvilinear trend is observed by the graph

Least square method

This method is totally mathematical and is free from all sorts of subjectivities. It is , a very appropriate and reliable method and is extensively practiced. The procedure I is exactly the same as described with fitting of regression line or a regression curve.

Fitting a line or a curve means estimating the parameters of the &. Equation and establish a prediction equation. The only difference between the regression line and a trend line is that in this, the X-variable is always the time period and the corresponding variate values stand for Y-variable.

Fitting of linear trend

Let us consider the linear trend as

$Y = \alpha + \beta x + e$

Where e $N(o, \varepsilon^2)$, Let a and b are the estimates of α and respectively. The estimated trend line is,

$Y_T = a + bX$

In the equation (1.4), a is the intercept and b is the slop of the line is

Under least square approach the quantity $\Sigma (Y - j)^2$ is minimized by the method of least squares. Also in this easel $\Sigma(Y - Y_T) = 0$ under the principle of least squares, we know,

$$b = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

The expression (1.6) is further simplified in this case by coding the middle year as zero and the years preceding it as -1,2,3, ... and following it as 1,2,3,.....if the number of years is odd. Also taking middle half year as zero and the year preceding this as — 1, -3, -5, . . . and following it 1,3,5,.....if the number of years in the series is even. This makes $\Sigma x = 0$ and the expression for b reduce to

$$b = \Sigma xy / \Sigma x^2 \qquad (1.7)$$

Substituting the values of a and b in the equation (1.4), we obtain the trend line.

Merits:

i. The least square method is devoid of all ambiguity and subjectivity.
ii. The estimates are unbiased and have minimum variance. Demerits:

Demerits

1) In the model it is assumed that Y depends on time alone which is not true in a large number of case. For instance, the increase in demand of various commodities is due to increasing population from year to year but not due to time factor. If the population growth rate is reduced the least square estimate will not be applicable.

Example 1-5:

The supply of electricity due to thermal projects from 1981 to 1989 ('000 M.W.) was as given below:

1981	1982	1983	1984	1985	1986	1987	1988	1989
17.6	19.3	21.4	24.4	27.0	30.0	31.8	35.6	39.7

The trend line by the method of least squares can be fitted by preparing the following computation table.

Year X	Coded (X) With 1985 = 0 X' - (X - 1985)	Y	X^2
1981	-4	17.6	16
1982	-3	19.3	09
1983	-2	21.4	04
1984	- 1	24.4	01
1985	0	27.0	00
1986	1	30.0	01
1987	2	31.8	04
1988	3	35.6	09
1989	4	29.7	16
Total	0	246.8	60

$$\sum x'y = 137, \sum x^2 = 60$$

$$a = \frac{246.8}{9} = 27.42$$

$$\text{and } b = \frac{137}{60} = 2.28$$

The trend line is,

$$Y_T = 27.42 + 2.28\,X$$

The trend values by putting the values of

X:-4, -3, -2, -1, 0, 1, 2, 3, 4 are,

1981	1982	1983	1984	1985	1986	1987	1988	1989
18.3	20.58	22.86	25.4	27.42	29.70	31.98	34.26	36.54

The estimated values are very near to actual values. Hence the trend line by the method of least squares is a goodfit.

Logarithmic trend line:

When the time series records per cent or proportional yearly changes, then a logarithmic linear trend is a better proposition. The equation of logarithmic trend is,

$$Y = a\,b^x \qquad (1.8)$$

If we take log of the equation (1.8) is,

$$\log Y = \log a - f \log b \qquad (1.9)$$

If we put log Y = z, log a = a_1 and log b = b_1

The equation (1.9) takes the form

$Z = a_1 + b_1X$ (1.9.1)

Which is a straight line equation (1.9) can be fitted by getting the values of log a and logb.

By the method of least squares,

$$\log a = \frac{\Sigma \log y}{n} \quad (1.10)$$

$$\log a = \frac{\Sigma x \log Y}{\Sigma x^2} \quad (1.11)$$

"Taking antilog we get the values of a and b respectively.

Parabolic trend: If the data shows a curvilinear trend, then a commonly faced curve is a second degree parabola. Its general equation is,

$Y = a + \beta x\ yx^2 + e$ (1.12)

Let the estimates of α, β an γd be a, b and c respectively. Using the method of least squares and replacing α, β an dγ by their estimates a, b and c, the normal equations are

$\Sigma Y = na + c\Sigma x^2$ (1.13)

$\Sigma xy = a\ \Sigma x + b\Sigma x^2 + c\Sigma x^3$ (1.14)

$\Sigma x^2Y = a\ \Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$ (1.15)

Under the system of coding as discussed in case of linear equation, sum of x of odd powers is zero. Therefore, Xi = Ix^3 = 0. Using these relation, the normal equations reduce to,

$\Sigma Y = na + c\Sigma x^2$ (1.16)

$\Sigma xY = b\ \Sigma x^2$ (1.17)

$\Sigma x^2Y = a\Sigma x^2 + c\Sigma x^2$ (1.18)

Solving the equation (1.16) through (1.18) for a, b and c we get the following, estimates,

$$a = \frac{\Sigma y - c\Sigma x^2}{n}, b = \frac{\Sigma xy}{\Sigma x^2}, c = \frac{n\Sigma x^2 y - \Sigma x^2 y - \Sigma x^2 \Sigma y}{n\Sigma x^4 - \left(\Sigma x^2\right)^2}$$

Substituting the estimated value for , α, β and γ and *y* we obtain the parabolic trend equation as,

$Y = a + bx + cx^2$ (1.19)

Parabola is not the only curvilinear trend but can be many more.

QUESTIONS

1 Following table gives the production of (thousand units) 1980 to 1991.

Years	Production (000 units)	Years
1980	405	1986
1981	415	1987
1982	450	1988
1983	465	1989
1984	470	1990
1985	472	1991

(i) Fit in a trend line by free hand

(ii) Fit in a trend line by semi-average method.

2 The sales of a firm as per records are:

Year	Sales ('000 Rs.)
1984	35
1985	38
1986	42
1987	37
1988	56
1989	48
1990	52
1991	30
1992	32

i. Fit the trend line by semi-average method.
ii. Find the trend of the data by 3-years moving average method.

3 Below are the figures of production (in thousand tonnes) of a sugar factory

Years	:	1981	1982	1983	1984	1985	1986	1987
Production	:	70	80	85	82	90	100	96

('000 Tonnes)

Fit in the line trend by least square method.

4 The following data relates to the number of scooters (in lakhs) sold by a manufacturing company during the Years 1985 to 1992.

Year	:	1985	1986	1987	1988	1989	1990	1991	1992
Number	:	6	6.1	5.2	5	4.6	4.8	4.1	6.2

(in Lakhs)

Fit a straight line trend and estimate the sales for the year 1993. (Take the year 1988 as working origin).

5 The following are data on the production (in '000 units) of a commodity from the years 1980 to 1986.

Year	:	1980	1981	1982	1983	1984	1985	1986
Production	:	6	7	5	4	6	7	8

('000 units)

Fit the trend of the type $Y = a + bx + cx^2$ to the above data. take 1983 as the year of origin).

6 Calculate four-yearly moving average from the following data

3Year	Annual values (in Rs. '000)
1960	52.7
1961	79.4
1962	76.3
1963	66.0
1964	68.6
1965	93.8
1966	104.7
1967	87.2
1968	79.3

7 Fit a linear trend equation to the following data by the method of least squares:

Year	:	1975	1976	1977	1978	1979
Production	:	83	92	71	90	169

'000 tonnes)

Estimate the production for 1980.

REFERENCES

Agarwal, B.L., *'Basic Statistics[1],* Wiley Eastern Ltd., New Delhi, 2nd ed, 1991.

Fuller, W.A., *'Introduction to Statistical' Time Series,* John Wiley, 1976.

Richard, L.E., Lacava, J., *'Business Statistics' (Why and When),* McGraw-Hill Book Company, 1978.

- End Of Chapter -

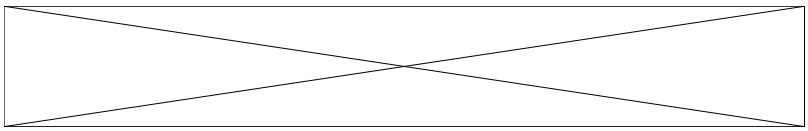
LESSON - 22

SEASONAL INDICES IN TIME SERIES ANALYSIS

Objectives

Seasonal variations occurring in a time series refer to those changes which occur during the periods (seasons) within a year. It gives an idea about the sales of woolens (winters). Similarly one can make an estimate of the sales of ice cream during summers. Seasons have lot of impact on the sales and hence unless an executive knows about the quantum of sales or nature of seasonal variations, we cannot take policy decisions. The objectives of study of seasonal variations is twofold.

i. To isolate the effect of seasonality in order to evaluate the effect of seasonal factors on a time series.
ii. To make the time series free from the effect of seasonal variation. This operation facilitates to make long term forecasts. The elimination of seasonal effects from a time series is known as *deseasonalisation.*



Please use headphones

To isolate the seasonal variations, one will have to remove trend cyclical and irregular variations. For a multiplicative model,

$$S = \frac{T \times S \times C \times I}{T \times C \times I} = \frac{Y}{T \times C \times I} \qquad (2.1)$$

and in case of an additive model,

$$S = Y - T - C - I \qquad (2.2)$$

The study of seasonal variations refers to two type of seasonals. They are:

(i) The season within a year is known as *specific seasonal*

(ii) On the other hand the average of specific seasonality over a number of years is known as *typical seasonal.*

Also a season in a year is not a bunch of any consecutive month but it refers to those in which really a particular season occurs, for instance rainy season occur from July to Sept. and not otherwise.

There are various method of calculating seasonal indices which are named below and then explained.

(a) Method of simple averages

(b) Ratio to trend method

(c) Methods based on moving averages

(d) Link relative method

(a) Simple average method :

The seasonal indices by this method are calculated in four steps.

(i) Arrange for each year according to the quarters, months etc. for which the seasonal indices are to be calculated.

(ii) Calculate the average for each month or quarter.

(iii) Compute the overall average.

(iv) Obtain the seasonal indices in percentages by dividing every month's averages by the overall average and multiplied by 100 one by one.

Check:

The exactness of seasonal indices can be checked by taking the sum of the indices. For monthly data, it will be 1200 and for quarterly data 400.

The simple average method is based on the assumptions:

(i) The upswing and downswing of cycles in the series are fairly balanced.

(ii) If any irregular movement are present, they are of random nature and compensate each other under averaging.

But such assumptions do not strictly hold and hence this method provides only rough estimates.

Example 2-1 :

The monthly production of fertilizer of a factory for four years was as given .below:

Month	**Production in 1akh (Tonnes)**				**Total**
	1987	1988	1989	1990	
Jan.	11	12	15	14	52
Feb.	12	14	16	18	60
Mar.	19	20	13	17	69
Apr.	13	17	16	19	65
May.	14	15	12	14	55
Jun.	12	13	9	10	44
Jul.	15	10	12	13	50
Aug.	20	18	16	22	76
Sep.	12	15	11	14	52
Oct.	14	10	12	13	49
Nov	13	14	15	16	58
Dec.	16	13	14	15	58
Total					688
Average					57.33

The monthly seasonal indices can be calculated in the following manner. Calculated values are given in the last three columns first to save space.

$$\text{Average of Jan.} = \frac{11+12+15+14}{4} = \frac{52}{4} = 13.0$$

$$\text{Average of Feb.} = \frac{12+14+16+18}{4} = \frac{60}{14} = 15.0$$

and so on

$$\text{Over all average} = \frac{172}{12} = \frac{57.33}{4} = 14.33$$

(Monthly) seasonal indices for,

$$\text{Jan.} = \frac{13.00}{14.33} \times 100 = 90.72; \quad \text{Feb.} = \frac{15}{14.33} \times 100 = 104.68$$

$$\text{Mar.} = \frac{17.25}{14.33} \times 100 = 120.38, \quad \text{Apr} = 113.34$$

and so on

Check :

The sum of the seasonal indices is little more than 1200. For a greater accuracy the indices can be adjusted by multiplying each seasonal index by the quantity 1200/(sum of indices).

Example 2-2:

Assuming that trend is absent in the data tabulated below,

Year	1st Quarter	2, Quarter	3 Quarter	4th Quarter
1989	3.9	5.0	3.6	3.2
1990	3.7	4.6	4.1	3.7
1991	4.5	4.2	4.5	3.5
1992	4.2	4.8	3.9	3.6

Calculate the seasonal indices for the four quarters.

Years	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1989	3.9	5.0	3.6	3.2
1990	3.7	4.6	4.1	3.7
1991	4.5	4.2	4.5	3.5
1992	4.2	4.8	3.9	3.6
Total	16.3	18.6	16.1	14.0
Average	4.075	4.650	4.025	3.5
Seasonal indices	100.31	114.46	99.08	86.15

1st Quarter Avg. $= \frac{16.3}{4} = 4.075$; 2nd Quarter Avg. $= \frac{18.6}{4} = 4.65$

3rd Quarter Avg. $= \frac{16.1}{4} = 4.025$; 4th Quarter Avg. $= \frac{14.0}{4} = 3.5$

Over all averages $= \frac{4.075 + 4.650 + 4.025 + 3.5000}{4} = 4.0625$

Check : The sum of the indices is 400

(b) Ratio to trend method :

This is also called percentage to trend method. This method is better than simple average method in the sense that it does not assume that seasonal variation for any month or quarter is a content factor in trend. Seasonal indices by ratio to trend method involves the following steps:

i. Obtain the trend value by the method of least squares for and appropriate trend equation (Usually a linear trend).
ii. Divide the original observation by this (estimated) trend value and multiply by 100. Each of these value is as the percentage of trend values. Assuming of the multiplicative model, the other steps are almost similar to simple average method.

iii. Find the average of the percentages for each quarter or month as the case may be. This eliminates the cyclic and irregular variations. These averages are the seasonal indices.

In some situation data possess extreme values. In that case median should be found out instead of average.

iv. If the indices for the months or quarters do not sum up to 1200 or 400 as per the situation, they can be adjusted by multiplying each indices by the factor 1200/sum of monthly indices or 400/sum of the quarterly, indices respectively.

Example 2-3 :

The seasonal indices for the following data.

Years	1st	2nd	3rd	4th
1985	5.5	10.0	6.5	10.0
1986	20.2	17.9	6.3	7.6
1987	9.5	17.2	10.6	10.7
1988	12.6	180	13.3	12.1
1989	14.4	16.8	5.2	13.6

Can be found out in the following manner.

First we calculate the trend values.

Year	coded (X)	Quarterly average for the year (Y)	XY	X^2	Trend values
1985	-2	32/4 = 8	-16	4	1010
1986	-1	54/4 = 13.5	-13.5	1	H.05
1987	0	48/4 = 12	00	0	12.00
1988	1	56/4 = 14	14	1	12.95
1989	2	50/4=12.5	25	4	13.90
Total		60	9.5	10	

Now we fit the trend line Yt = a + bx

$$\text{Where, } a = \frac{\Sigma y}{n} - = \frac{60}{5} = 12.0$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{9.5}{10} = 0.95$$

Thus, the trend line is

Yt = 12 + 0.95 x

For this equation 1987 is origin 0 and

Yearly increment is 0.95

Hence, the quarterly increment = 0.95/4 = 0.24

Let us call the quarterly increment 'C = 0.24

The trend values have been calculated by putting x equals to -2, -1, 0 and 1, 2 respectively and entered in the last column of the above table. The trend value lies at the mid year. So the trend value for the second quarter will be trend value -C/2 and for the 3rd quarter as trend value -f c/2. Similarly for 1st and 4th quarter will be, trend value-3C/2 and trend value +3c/2. Using this adjustment we write the trend values for each quarter and for all the years as follows:

	Quarterly Trend varieties			
Years	**1st**	**2nd**	**3rd**	**4th**
1985	9.74	9.98	10.22	10.46
1986	10.69	10.93	11.17	11.41
1987	11.64	11.88	12.12	12.36
1988	12.59	12.83	13.07	13.31
1989	13.54	13.78	14.02	14.26

For 1985, $X = -2$ yearly trend value $= 12 - 0.95 \times 2 = 10.10$

Year 1985, for 1st quarter trend value $= 10.10 - \frac{3 \times .24}{2}$

$= 10.10 - 0.36 = 9.74$

Year 1985, for 2nd quarter trend value $= 10.10 - \frac{.24}{2} = 9.98$

Year 1985, for 3rd quarter trend value $= 10.10 + \frac{.24}{2} = 10.22$

Year 1985, for 4th quarter trend value $= 10.10 + \frac{3 \times .24}{2}$

Trend $= 10.46$.

Similarly, the quarterly trend values for the other years have been calculated and entered in the above table.

For 1986, $x = 1$, the yearly trend value $= 12 - 0.95 = 11.05$.

1st quarter, trend value $= 11.05 - \frac{3 \times .24}{2} = 10.69$

and so on.

and so on.

Now we calculate trend eliminated values which are obtained dividing the quarterly value by the corresponding trend value multiplied by 100. These as displayed in the table below:

Year	Trend eliminated value Quarterly values per centage tr dues		
	1st quarter	**2nd quarter**	**3**
1985	56.47	100.20	63.6
1986	188.96	163.77	56.4
1987	81.61	144.78	87.4
1988	100.08	140.30	101.
1989	106.35	121.92	37.0
Total	533.47	670.97	436.
Average Seasonal indices	106.69	134.19	69.2
Adjusted seasonal indices	107.46	135.15	69.7

The sum of the seasonal indices - 397.15. Now to bring the sum to 400 each seasonal index is multiplied by 400/397.15 and entered in the last low. It can be verified that the sum of the adjusted seasonal indices is 400.

Note: The same method can be applied to monthly seasonal indices. In this-situation, year wise monthly data have to be treated in the same way as we did for a quarter.

Ratio to moving average method :

Just like trend, moving average method is very popular for finding out the seasonal indices. The main advantage of this; method is that it eliminates periodic changes if the period of moving average is equal to the cycles and are to be eliminated. In moving average method we take, 12-months or 4-quarters moving averages for the monthly or quarterly data respectively. Thus, this eliminates the seasonal variation totally provided they are constant in their amplitude and direction. The method involves the following steps.

Step-1 : Write the monthly or quarterly data chronologically.

Step-2: Find the twelve-monthly or 4-quarters average for the first year and enter it in the mid-position of the 12-months (between June and July) or of 4-quarter in the middle of II and III quarter.

Step-3 : Delete the January value for the first year and add the January value for the next year and again find the average and enter it between July and August. In case of quarterly date, delete first quarter value and add next years quarter value.

Find the average of this new set of values and enter it against the III and IV quarter. Continue this process until all the monthly or quarterly data are exhausted.

Step-4 : Since, the averages under step-3 are entered against any month or quarter, again find the moving averages of the two average obtained under step-3 and enter it in front of the July or III quarter according to the situation.

Step-5 : Calculate the ratio of each monthly (quarterly) value to the corresponding moving average value multiplied by 100. Obtained under step-4 and enter it against the month it exists.

Step-6 : Now prepare two way table displaying the years and monthly (quarterly) ratio obtained under step-5.

Step-7: Find the media for each month or quarter. These medians are nothing but seasonal indices

Step-8: If the sum of these indices is not 1200 or 400 in case of monthly or quarterly data respectively, then it should be adjusted by multiplying each index by 1200/sum of indices.

Remark :

The above method is suitable if the multiplicative model is used. In practice, mostly the multiplicative model is used.

In case of additive model instead of percentage ratio to moving average, we have to compute deviation from moving averages.

Example 2-4 :

Given the following quarterly data for four years, calculate the seasonal indices by the method of moving averages using multiplicative as well as additive model.

Years	Quarters	Y	4 - Quarters moving averages	Moving average o two averages
	I	27	-	-
	II	23	-	.
			28.50	
1988	III	35		28.125
			27.75	
	IV	24		27.125
			26.50	
	I	24		25.625
			24.75	
	II	23		26.00
			27.25	
1989	III	28		28.875
			30.50	
	IV	34		32.000
			33.50	
	I	37		35.625
			37.75	
	TT	32:		36.750
	XL		35.75	
1990	III	42		35.625
			35.50	
	IV	26		35.625
			35.75	
	I	36		36.000

Assuming multiplicative model, the ratio to moving averages and then seasonal indices are calculated and displayed in the following table. The values in the body of the following table represent the percentage ratios as (Y/corresponding moving average) x 100.

		Quarters		
Year	I	II	III	IV
1988	__	__	124.44	88.48
1989	93.66	88,46	96.97	106.25
1990	103.86	103.40	117.89	72.98
1991	100.00	102.63	-	-
Medians (seasonal)	100.00	102.63	117.89	88.48
Adj. seasonal indices	97.80	100.37	115.30	86.53

The sum of the seasonal indices = 409.00. The adjusted seasonal indices are entered in the last row which are obtained by multiplying each index by the fraction (400/409). After adjustment, the sum of the indices reduces to 400.

If the additive model is assumed, the trend eliminated values are entered in the following table and the seasonal indices for the quarters are calculated by taking the average of the trend eliminated values.

		Quarters		
Year	**I**	**II**	**III**	**IV**
1988	_		6.875	-3.125
1989	-1.625	-3.000	-0.875	2.000
1990	1.375	1.25	6.375	-0.625
1991	00	1.00	-	-
Total	0.250	-0.750	12.375	-10.75
Average (seasonal)	0.083	-0.250	4.125	-3.583
Adj. seasonal indices	-0.01075	-0.34375	4.03125	-3.67675

Sum of the quarterly indices = 0.375

$$\text{Adjustment factor} = = \frac{1}{4} \times 0.375 = 0.9375$$

To obtain the adjusted seasonal indices, subtract the adjustment factor C = 0.09375 from each of the quarterly average. The same values are entered in the last row of the above table.

Remark :

The same procedure can be adopted for monthly data and monthly seasonals can be obtained.

Link relative method:

This method was invented by Karl Pearson. Link relative method for finding out the seasonal indices has been explained through the following steps.

Step-1: Express each periodic or seasonal value as the percentage of the preceding value of the time series. The percentage so obtained represents the *link relatives* (L.R.). This eliminates the influence of the trend. Besides the trend the cyclic effects are also eliminated to a great extent.

Step-2 : Calculate the median of the link relative for each period or seasons (months or quarter etc.) This eliminates irregular effect. These medians are not seasonal indices but only the medians of the link relative.

Step-3 : Calculate *chain relative* (C.R.) *medians* on the basis of the first season, month or quarter etc. obtain other season's chain relative medians. Assume first season median L.R. as 100. The formula for chain relatives is,

$$= \frac{\text{Median for the period} \times \text{previous seasons C.R}}{100}$$

Step-4 : Calculate the chain relative for the first season on the basis of the last season. First season chain relative is

$$= \frac{\text{Median L.R for the I seasons} \times \text{C.R median for the last season}}{100}$$

Since the variation present due to long distance between first season and last C.R. some error is induced which needs correction. Hence, a correction factor is introduced which is,expressed in Step-5.

Step-5 : The adjusted value of chain relative median for the first season is taken equal to 100. For this, the adjustment factor C is obtained by the formula,

$$C = \frac{100 - \text{C.R. for the first season}}{\text{number of season}}$$

Number of seasons for monthly data is 12, Hence the correction factors for the seasons from first to last one are OxC, IxC, 2xC, respectively. The correction factors are added to the season's chain medians.

Step-6 : Find the mean of the adjusted medians, Divide each seasons median by the mean of the medians. The resulting values represent the seasonal indices. The advantage of this operation is that the sum of the final indices reduces to 1200 or 400 etc. as the case be:

Advantages of link relative method

1. The link relatives eliminate the cyclic and trend effects. ,
2. Link relative are calculated for each and every season. So no information is lost.
3. The use of correction factor further eliminates the trend effect.

Example 2-5 :

Compute seasonal indices for the following quarterly data by the method of link relatives.

			Years		
Quarters	**1984**	**1985**	**1986**	**1987**	**1988**
I	121	128	134	153	162
II	115	130	129	147	165
III	133	134	131	169	195
IV	169	125	151	154	217

Firstly we calculate the link relative and enter the following table.

For the 1984, 1st link relative $= \frac{115}{121} \times 100 = 95.04$

For the 1984, 2nd link relative $= \frac{133}{115} \times 100 = 115.65$

Year	**I**	**II**	**III**
1984	127.16	95.04	115
1985	75.74	101.56	103
1986	107.20	96.27	101
1987	101.32	96.07	111
1988	105.19	101.85	118
Median	103.26	96.27	114
Chain relative median	100	96.27	110
Adjusted chain relative median	100	89.48	97.
Seasonal indices	102.73	91.92	99.

Chain relative median is calculated in the following manner.

Chain relative for I quarter $= 100$

Chain relative for II quarter $= \frac{96.27 \times 100}{100} = 96.27$

Chain relative for III quarter $= \frac{114.96 \times 96.27}{100} = 110.67$

Chain relative for IV quarter $= \frac{103.26 \times 110.67}{100} = 123.15$

The Chain relative for the I quarter $= \frac{103.26 \times 123.15}{100} = 127.16$

The first chain relative is shown in the box in position $(1,1)$.

The correction factor, c $= \frac{100 - 127.16}{4} = 6.79$

The adjustment factor for I quarter $= 0 \times -6.79 = 0$

The adjustment factor for II quarter $= 1 \times -6.79 = -6.79$

The adjustment factor for III quarter $= 2 \times -6.79 = -13.58$

The adjustment factor for IV quarter $= 3 \times -6.79 = -20.37$

The average of the adjustment chain relative mwdian is

$$= \frac{100 + 89.48 + 97.07 + 102.78}{4} = 97.34$$

The final seasonal indices for

Ist quarter $= \frac{100}{97.34} \times 100 = 102.73$

II quarter $= \frac{89.48}{97.34} \times 100 = 91.92$

III quarter $= \frac{99.09}{97.34} \times 100 = 99.74$

IV quarter $= \frac{102.78}{97.34} \times 100 = 105.59$

The sum of the seasonal indices is 399.98 which is almost 400.

Isolation of cyclic and irregular variations

Once we remove the trend and seasonality from the data, it is left with cyclic and irregular factors. Removing seasonal effect from the data is known as deseasonalisation.

There are various methods of isolating cyclic and irregular variation but they are not discussed as they are not the part of the syllabus.

QUESTIONS

l The prices of manufactured products from the year 1980 to 1986 are as follows:

Years :	1980	1981	1982	1983	1984	1985	1986
Prices : (Crore Rs.)	108	112	114	119	114	125	120

Fit in the trend line by (i) graphical method (ii) least square method.

2 The following tables gives the number of workers employed in a small industry during the years 1979-88. Calculate the four yearly moving average.

Year:	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
No. of : Workers	430	470	450	460	480	470	470	500	430	480

3 Explain briefly, how the seasonal element in a time series data is isolated and eliminated.

4 The number (in hundreds) of letters posted in a certain city on each day in a, typical period of five weeks was as follows:

Week	Sun.	mon	Tues.	Wed.	Thurs,	Fri	Sat
1st	18	161	170	164	153	181	76
2nd	18	165	179	157	168	195	85
3rd	21	162	169	153	139	185	82
4th	24	171	182	170	162	179	95
5th	27	162	186	170	.170	182	120

Obtain the indices of seasonal variation, within the weeks, by the simple average method.

Construct the index of seasonal variations from the following data:

Year	Jan.	Ml	Mat	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov	Dec.
1980	18	20	18	17	15	16	17	19	19	23	23	24
1981	20	22	10	18	IS	20	24	23	23	24	24	26
1982	22	18	20	18	17	18	24	25	26	24	25	27
1983	24	24	22	20	18	22	25	26	27	26	27	29

By the method of (i) 4 months moving average method (ii) ratio to trend method (iii) Link relative method for multiplicative model.

6 Calculate seasonal indices by the ratio to moving average method from the following data:

Year	I	Quarter II	III	IV
1981	68	62	61	63
1982	65	58	66	61
1983	68	63	63	67

Q.7 Deseasonalise the following data with the help of a suitable method:

Month	Cash Balance ('000 Rs)	Seasonal . Index
January	360	120
February	400	80
March	550	110
April	360	90
May	350	70
June	550	100

8 The data given below gives the average quarterly prices of a commodity for four years:

- End Of Chapter -

LESSON - 23

STATISTICAL QUALITY CONTROL

Introduction

In nature, no two things are similar as Bernard Shaw wrote: "Nature never repeats". But in a manufacturing process, it is expected that all items will be alike but the experience is dispelled this belief. If we take measurements on the manufactured item of a factory it has been found that they too differ. If the difference is minute, it can be neglected and if large, it has to be taken care off. So, there was a great problem whether to accept a lot for marketing or not. If the product does not meet the specifications and marketed, it will affect the reputation of the company. Again if the lot is rejected even for slight variation from specifications, the company might have to undergo unbearable losses. So the statistical quality control comes to the rescue of the manufacturer. Under this, statistically two limits are set, the lower and upper limits for a particular variety and if all items measured lie within these limits, the

process is said to be under control and if not, the process is considered to be out of control. Hence, the fault has to be removed and a particular lot may be rejected.

Objectives

Statistical quality control meets the following objectives in a manufacturing process.

i. To have a constant vigil whether the process is under control or not.
ii. To have a check on quality of the product at a low cost.
iii. To have a timely check on the process,
iv. To provide mathematical basis for the acceptance or rejection of a lot.
v. To have a check on the quality of items which are destroyed under the process of testing.
vi. To ascertain the impact of change of persons or process on the quality of the product.

Causes for v ariation

Two types of causes are responsible for variation.

i. Chance causes
ii. Assignable causes

i. Chance causes: Some variation in the manufactured product or articles is bound to occur, howsoever, the process may be efficient. Now. If this variation occurs due to certain inherent pattern of variability. It is considered to be the variation due to chance causes. Such a variation cannot be controlled. The product is considered to be of good quality and the process is considered to be under control.
ii. Assignable causes: If a product shows a marked variation in the specifications, then the product cannot be considered of good* quality. In this situation, it is expected that there are some faults in the manufacturing process which are responsible for such a large variation and can be amended. Such causes which are detectable and removable, are known as assignable causes.

Role of statistical quality control

Statistical quality control is meant to collect data from the manufacturing process at various stage or at the final stage and analyse the data for the purpose of retaining whether the process is under control or not. If not, what are possibly assignable causes. Mostly statistical quality control methods are base control charts.

Control Charts :

Control charts are the devices to show .the pattern of variation in the units inspected. Control charts delimit the range or band in which the variability of a product is under tolerance limits and outside these. Limits the variability is not tolerable. The most frequently used control charts are, the mean X chart, the range R chart, standard deviation a chart and the number of defects C charts. Control charts leads are to

conclude whether the process is under control or not. If not, what are the assignable causes for it. These defects or faults can be removed and the manufacturing process is brought under control.

Rationale behind the control charts :

We know that for a normal population distributed with mean and standard deviation , 99.73 : units lie within the limits pi - 3a and pi+3a. If the population parameters fj, and a are not known their estimated values, sample mean X and sample standard deviation o can be used. The process is considered under control if the observed relevant value lie within their limits. If the observed values are outside their limits, the process is said to be out of control.

A control chart is essentially a graphic method for presenting data so as to reveal the extent of variation from the limits at a glance. A control chart contains three line, namely:

(i) the control line, which shows the standard which one wants to maintain in respect of certain characteristic of a product;

(ii) lower control limit, (iii) Upper control limit. These limits provide a band, in which the values shows that the process is under control limits and any point lying beyond these limits show that the process is not under control. A sketch of the control chart is given below in fig.(3.1).

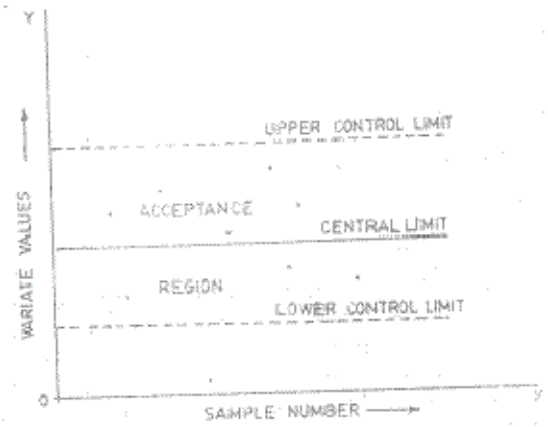


Fig.(23.1) Control chart in general

Fig.(23.1) Control chart in general

Two types of control charts

Control charts for statistical quality control are classified into two types:

(i) control charts for variables

(ii) control charges for attributes.

Control chart for variable is prepared when the variable is continuous. In this situation, the mean, standard deviation and range are found out for the samples drawn at regular interval of production process and thus X a and R charts are prepared.

Control charts for attributes are prepared in that situation where the sampled units are inspected for finding out whether an unit is defective or non defective. Also many times it has been checked how many defects per unit are there in this situation d or C charts are usually prepared.

Interpretation of Control Charts

If any point out of the plotted points lies outside the control limits, then the process is considered to be out of control and hence some assignable causes be traced out and necessary corrective measures be taken to bring the process under control. This saves the manufacturer from heavy losses. Again if all the points (dc^f.s) lie within the control limits the process is considered to be under control and whatever variation is being observed is due to chance (random) causes.

Sometime, the points (dots) may be lying within the control limits but still they show a peculiar pattern. For example all the points may be lying below the central line or above it or they follow a peculiar path. All such arrangement should not be left unnoticed. They are also the danger signals which may indicate a change in the production process. Hence, the control charts should not be scrutinized only for the dots lying within the control limits or outside but should also be taken care for some special pattern. For the first time, the control charts were developing by Dr.Walter A.Shewhart of the Bell Telephone Company in 1924, since then they are in consistent use. Mainly, X, , R and C charts are used which are discussed in succession.

Setting of control limits :

To construct the control limits usually one draws K samples of equal size n at regular intervals from the manufacturing process and take observations on them. Let observations for the ith sample be denoted by Xi_1, Xi2,....., Xin for i = 1,2, ... n.

As already mentioned 99.73% units lie within the limits u ± 3 a in case of normal populations. But in a large number of cases, the mean JU and standard deviation a are not known. In this case, the estimates are used. We know the standard error of a sample mean is $\frac{s}{\sqrt{n}}$ where s is the sample standard deviation.

$$X_i = \frac{I}{n} \sum_{j=1}^{n} x_{ij} \qquad (3.1)$$

and the overall mean (mean of the means) is,

$$X_i = \frac{I}{k} \sum_{j=1}^{k} x_i \qquad (3.2)$$

The variance of the 1th sample

$$S_i^2 = \sum_{j=1}^{n} \left(x_{ij} - x_i\right)^2 \qquad (3.3)$$

$$S_i = \sqrt{s_i^2} \qquad (3.4)$$

The estimate of a is,

$$S = \frac{1}{k} \sum_{i=1}^{k} 8_i \qquad (3.5)$$

The control limits will be mean ± 3(S.D. of the mean)

X-Chart

X is the over all sample mean and is the estimate of population mean ft .

An unbiased estimate of the standard deviation a $\frac{\bar{s}}{C_2}$ where $C_2 = \sqrt{\frac{2}{n}} \frac{\sqrt{\frac{n}{2}}}{\sqrt{\frac{n-1}{2}}}$

control limits are

$$X \pm 3 \frac{\bar{s}}{C_2 \sqrt{n}} \qquad (3.6)$$

Thus the lower control limit (L.C.L.), central limit (C.L.) and the upper control limit (U.C.L.) are,

$$\begin{aligned} L.C.L &= X - 3\frac{\bar{s}}{C_2\sqrt{n}} \\ C.L &= X \qquad (3.7) \\ U.C.L &= X + 3\frac{\bar{s}}{C_2\sqrt{n}} \end{aligned}$$

Now Putting A, For $\frac{3}{C_2\sqrt{n}}$

The control limits for X chart can be written as

$$\begin{aligned} L.C.L &= X - A_1\bar{S} \\ C.L &= X \qquad (3.7) \end{aligned}$$

In case the specification for the mean and standard deviation are already set as ju and a respectively, then the control limits of X — chart will be

$$\begin{aligned} L.C.L &= \mu^1 - A\sigma^1 \\ C.L &= \mu^1 \qquad (3.8) \\ U.C.L &= \mu^1 - A\sigma^1 \end{aligned}$$

σ-Chart

Many scientist prefer standard deviation as equality characteristic instead of mean. Let V be the sample standard deviation. Then by definition, the variance of x can be given as

$$V(s) = E(s^2) - [E(s)]^2 \qquad (3.9)$$

We know

$$E(s^2) = \frac{n-1}{n}\sigma^2 \text{ and } E(s) = C_2\sigma$$

$$\therefore V(s) = \frac{n-1}{n}\sigma^2 - C_2^2\sigma^2 \qquad (3.10)$$

$$= \left(\frac{n-1}{n} - C_2^2\right)\sigma^2 \qquad (3.10.1)$$

$$= C_3^2\,\sigma^2 \qquad (3.11)$$

$$\text{Where } \frac{n-1}{n} - C_2^2 = C_3^2$$

$$\therefore s.D.(s) = C_3\sigma$$

Thus the contorl limits for σ - chart are,

$$L.C.L.s = E(s) - 3.sE.(s) \qquad (3.12)$$

$$= C_2\sigma - 3C_3\,\sigma$$

$$= (C_2 - 3C_3)\,\sigma$$

$$C.L.s = C_2\sigma$$

$$U.C.L.s = (C_2 + 3C_3)\,\sigma$$

If we put C2 - 3 C3 = Bi and C2 + 3C3 = B2, then the control limits given in (3.12) is reduce to

$$L.C.L.s = B_1\sigma$$

$$C.L.s = C_2\sigma \qquad (3.13)$$

$$U.C.L.s = B_2\sigma$$

When σ is not known, its estimated value is used which is equal to $\frac{\bar{s}}{C_2}$

Hence, the control limits when a is not known are given as follows:

$$\text{L.C.L.}_s = (C_2 - 3C_3)\frac{\bar{s}}{C_2} = \left(1 - \frac{3C_3}{C_2}\right)\bar{s}$$

$$\text{C.L.}_s = C_2\frac{\bar{S}}{C_2} = \bar{s} \qquad (3.13)$$

$$\text{U.C.L.}_s = (C_2 - 3C_3)\frac{\bar{s}}{C_2} = \left(1 + \frac{3C_3}{C_2}\right)\bar{s}$$

If we put C2 - 3C3 = B3 and C2 + 3C3 = B4, the control limits given in (3.14) can be rewritten as,

$$\text{L.C.L.}_s = B_3\bar{s}$$

$$\text{C.L.}_s = \bar{s} \qquad (3.15)$$

$$\text{U.C.L.}_s = B_4\bar{s}$$

Remark :

If the lower control limit comes out to be negative it has to be taken as zero, since the standard deviation can never be negative.

R-chart

It has been experienced that in case of small samples, the standard deviation and the range fluctuate simultaneously. Also in general only small samples are drawn in case of quality control inspection. Hence, one can use the range in place of standard deviation to set up the control limits. The reason being that the computation of range is much easier than the computation of standard deviation. Therefore, the use of range is preferable as compared to standard deviation even for a little loss of efficiency. The relations between the range R and standard deviation a from the sampling distribution of range are given as

$$E(R) = d_2\sigma \qquad (3.16)$$

$$\text{S.D.}(R) = d_3\sigma$$

If there are K samples and Ri is the range of the i sample for i = 1, 2, . . . , k and H is the mean range of all the samples

Where $\bar{R} = \frac{1}{k} - \sum_{i=1}^{k} R_i$, Then $E(R_i) = \bar{R}$ Therefore

$\bar{R}$ $d_2\sigma$ or $\sigma\frac{\bar{R}}{d_2}$ $\bar{R}$ (3.17)

Case (1) : when the range R and its standard deviation a R are the known values of range R and S.D. o R , the control limits for R-chart are given as,

$$\begin{aligned} L.C.L_{R1} &= E(R) - 3S.D(R) \\ &= d_2\sigma R_1 - 3d_3\sigma R_1 \\ &= (d_2 - 3d_3)\sigma R^1 \qquad (3.18) \\ C.L.R^1 &= E(R) = d_2\sigma R_1 \end{aligned}$$

Similarly,

$$U.C.L.R^1 = (d_2 + 3d_3)\sigma R^1$$

If we put $d_2 - 3d_3 = D_1$ and $d_2 + 3d_3 = D_2$,

The control limits for R – chart are,

$$\begin{aligned} L.C.L.R_1 &= D_1\sigma R^1 \\ C.L.R^1 &= d_2\sigma R^1 \qquad (3.19) \\ U.C.L.R^1 &= D_2\sigma R^1 \end{aligned}$$

Case (ii) : When the value of population range R is not known we make use of the estimated value of R which is 7? as already clarified.

For X-chart, the control limits are,

$$L.C.L._{\bar{x}} = \bar{\bar{X}} - 3\sigma_{\bar{x}}$$

$$= \bar{\bar{X}} - \frac{3}{d_2\sqrt{n}}\bar{R}$$

$$= \bar{\bar{X}} - A_2\bar{R} \qquad (3.20)$$

$$C.L._{\bar{X}} = \bar{\bar{X}}$$

$$U.C.L._{\bar{x}} = \bar{\bar{X}} + \frac{3}{d_2\sqrt{n}}\bar{R}$$

$$= \bar{\bar{X}} + A_2\bar{R}$$

For R-chart, the control limits are,

$$= d_2\frac{\bar{R}}{d_2} - 3d_3\frac{\bar{R}}{d_2} \qquad (3.21)$$

$$d_2\frac{\bar{R}}{d_2}$$

$$d_2\frac{\bar{R}}{d_2}$$

$$= \left(1 - \frac{3d_3}{d_2}\right)\bar{R}$$

$$C.L._R = d_2\sigma = d_2\frac{\bar{R}}{d_2} = \bar{R}$$

Similarly,

$$U.C.L._R = \left(1 + \frac{3d_3}{d_2}\right)\bar{R}$$

Now putting $1 - \frac{3d_3}{d_2} = D_3$ and $1 + \frac{3d_3}{d_2} = D4'$

The control limits for R – chart given in (3.12) can be rewritten as, Now

$$L.C.L._R = D_3\bar{R}$$

$$C.L._R = \bar{R}$$

$$U.C.L_{R} = D_4 \overline{R}$$

Remark : The values of the constants A, Ai, A2, C2, B], B2, B3, B4, d2, Di, D2, D3, and D4 can directly be obtained from the table of factors useful in construction of control charts given by shew Rant. The same tables are reproduced in the book 'Basic Statistics* by B.L.Agarwal, appendix table XV

Example 3.1 : The following table gives the height in grams of the packets sampled on 15 days. The packets were sampled for 25 grams each. Ever}[7] day a sample of 5 packets was selected at random and weighted. The weights were as tabulated below:

Sample No.	Obed /eights (iii)					Total Mean		Range	Standard-deviation
	(i)	(i)	(iii)	(iv)	(v)		X	R	
1.	27	20	31	2	2	12	25	11	4.18
2.	36	31	26	4	3	5	29	14	5.29
3.	25	16	15	2	3	14	19	10	3.94
4.	22	28	32	2	0	5	26	13	5.34
5.	30	29	27	1	2	95	30	6	2.24
6.	34	28	34	9	0	13	30	6	4.24
7.	21	27	23	2	1	0	22	8	3.16
8.	11	22	28	9	9	15	23	20	7.65
9.	26	35	32	3	3	0	30	9	3.67
10.	14	18	22	1	3	15	20	12	4.47
11.	26	32	32	2	3	0	29	7	3.32
12.	19	28	29	4	0	11	24	10	4.53
13.	12	18	22	1	2	0	19	16	6.24
14.	27	36	47	9	0	11	32	25	9.77
15.	11	21	20	3	2	5	18	11	4.53
				1	3	15			
				2	3	0			
				7	0	10			
				2	2	0			
				0	6	14			
				3	2	5			
				0	5	12			
				2	2	0			
				4	0	95			
				2	1	16			
				8	5	0			
				2	2	90			
				8	2				
				1	2				
				6	2				
Total						1880	376	178	72.57
Average						25.06	25.06	11.87	4.84

The process is under control or not has been checked by constructing (i) X - chart, (ii) R — chart and (iii) s— chart.

(i) The control limits for X -chart from (3.6) are,

$$\overline{X} = \frac{3\bar{s}}{C_2\sqrt{n}}$$

The calculated values of X, R and s are entered in the last column of the above table to save space. From the above table,

n=5. X= 25.06, $\bar{s}$ = 4.84

From the table XV of Basic Statistics by B.L.Agarwal the value of the constant C2 for n = 5 is 0.9490. Substituting the values of X ,S and C2 and n, the control limits are,

$$\text{L.C.L.}_{\bar{x}} = 25.06 - \frac{3\text{x}4.84}{0.9490\text{x}\sqrt{5}}$$

$$= 25.06 - \frac{14.52}{0.949\text{x}2.24}$$

$$= 25.06 - \frac{14.52}{0.949\text{x}2.24}$$

$$= 25.06 - 2.12$$

$$= 22.94$$

$$\text{C.L.}_{\bar{x}} = 25.06$$

$$\text{U.C.L.}_{\bar{x}} = 25.06 + 2.12$$

$$= 27.18$$

Now we mark the control limits on the graph paper. Also plot the mean values against the sample numbers. Now check whether any point lies outside the control limits or not.

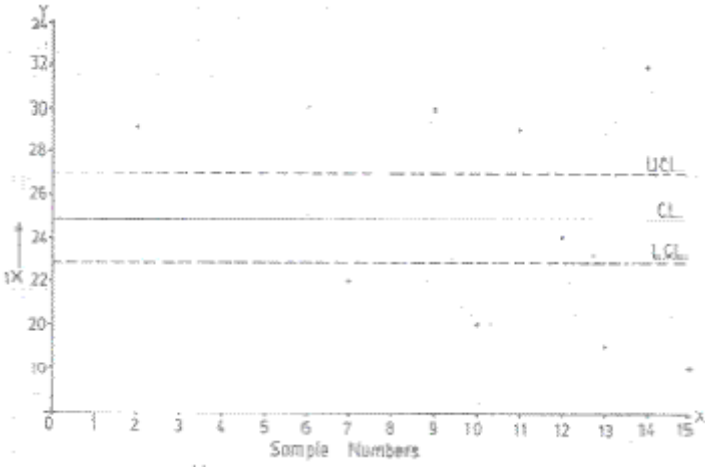


Fig. 23.2 X - Chart.

Since the points lie below the lower control line and also beyond the upper control line, we conclude that the process is not under control.

(ii) The control limits for the R-chart from (3.21.1) are,

L.C.L.r = $D3^R$

C.L.r = R

$U.L._R$ = $D4^R$

The value of 7t from the above table is 11.87 also from the table of constants, for n = 5, are D3 = 0 and D4= 2.115

Thus the control limits are,

L.C.L.r = 0x11.87 = 0

C.L.r = 11.87

$U.L._R$ = 2.115x11.87

= 25.10

Now we mark the control limits on the graph paper. Plot the values of range against sample numbers. The graph is as given below.

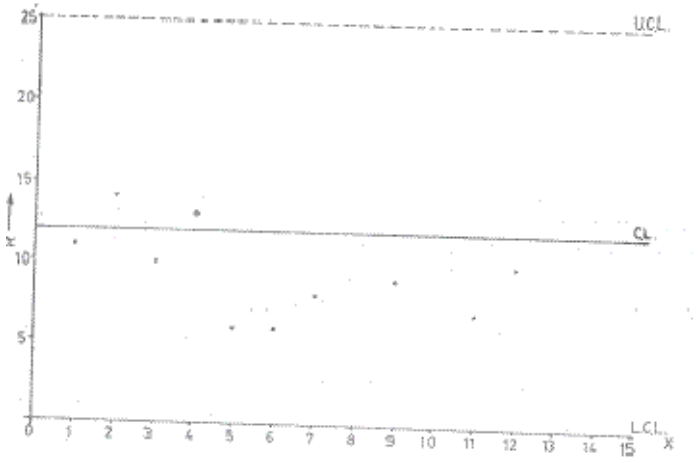


Fig. 23.3 R- chart

Since the plotted points lie above the upper control limit, we conclude that the process is not under control.

(iii) When the population value of the standard deviation are not known, the control limits are given by (3.15)

L.C.L.R = $B3\bar{s}$

C.L.R = $\bar{s}$

U.L.R = $B4\bar{s}$

From the above tableS = 4.84 and from the table of constants for n= 5, B3= 0 and B4 = 2.089. Substituting the values of B3, B4 and S, the control limits are,

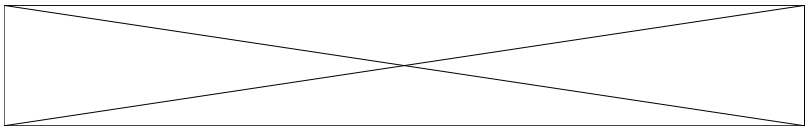
L.CL.s = Ox 4.84

= 0

C.L.s = 4.84

UC.L.s = 2.089 x 4.84

= 11.11



Please use headphones

Now we mark the control limits on the graph paper and also plot the values of the standard deviations against sample numbers as shown in the graph (3.4).

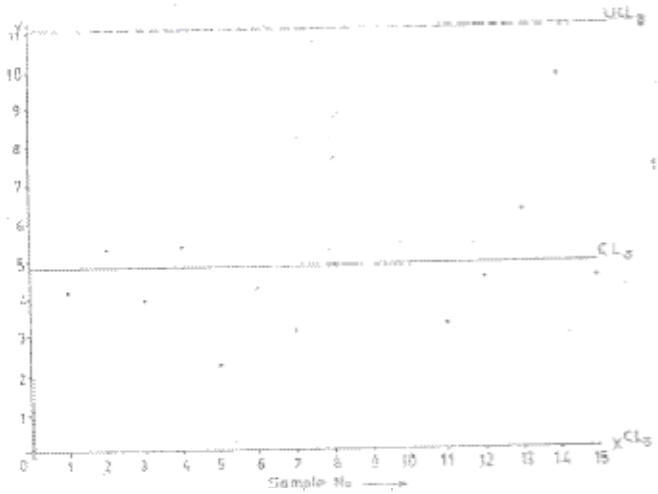


Fig. 23.4 a - Chart

After a check of the graph, we find that no point lies beyond the upper control limits. Hence, we come to conclusion that the process is under control in respect of standard deviation as the quality characteristics.

Example 3-2 : The following data show the values of sample mean X and range R for the samples of size 5 each. Calculate the value for central line and control limits for mean chart and range chart and determine whether the process is in control.

Sample No :	1	2	3	4	5	6	7	8	9	10
Mean(X) :	11.2	11.8	10.8	11.6	11.0	9.6	10.4	9.6	10.6	10.0
Range (R) :	7	4	8	5	7	4	8	4	7	9

(conversion factors for n = 5 and A2 = 0.577, D3 = 0 and D4 = 2.115)

Solution:

Over all mean $\overline{\overline{X}} = \frac{106.6}{10} = 10.66$

Mean range $\overline{R} = \frac{63}{10} = 6.3$

Control limits for the mean chart from (3.20) are,

C.L $= X + A_2R$

C.L.s $= X$

U.C.L. $= X + A_2R$

Substituting the values of X , R HandA_2weget

L.C.L.X = 10.66 - 0.577 x 6.3

=10.66 - 3.63

=7.03

U.C.L.X = 10.66 + 3.63

=14.29

control limits for range chart from (3.22) are,

L.C.Lr = D3K

0x6.3 = 0 C.L.r = 6.3

U.C.L.r = 2.115x6.3

=13.32

None of the sample mean lies outside the limits 7.03 and 14.29. Hence, the X-chart reveals that the process is under control. Again all the sample ranges lie within the control limits of 0 to 13.32. Hence, we conclude that the process is under control.

Control charts for attributes

In the beginning it has been given that there are charts for variables and the other for attributes. The charts for variables have already been discussed. Now we discuss the chart for attributes. This covers two types of charts:

(i) the charts for number of defectives, so called p-charts

(ii) the chart for number of defects per unit known as c-chart.

p-chart

If there are n units in the sample and d items are defective. Then, the proportion p of defective is d/n. Since, we are dealing with two types of items namely, defectives» and non defectives, the population is dichotomous. Hence, it can be thought of to follow binomial distribution. Using the mean and variance formula for binomial population, we can construct the control limits for p-chart. In the construction of control limits, we are using the following notations. If the proportion 'P' of the binomial is prefixed, then its standard

deviation is $\sqrt{P^1(1-P^1)/n}$

Thus, 3 a -limits for p-chart are,

$$L.C.L.p^1 = p^1 - 3\sqrt{\frac{p^1(1-p^1)}{n}}$$

$$C.L.p^1 = p^1 \qquad (3.22)$$

$$U.C.p^1 = P^1 + 3\sqrt{\frac{p^1(1-p^1)}{n}}$$

If the standard value of the proportion of defectives is not known, then its estimated value p has to be used. Suppose k samples each of size n are drawn at regular intervals from the manufactured items. Then, the estimated value,

$$\bar{p} = \frac{\text{Total number of defectives in all samples}}{n \times k} \qquad (3.23)$$

3 σ control limits for proportion of defectives are,

$$L.C.L.\bar{p} = p^1 - 3\sqrt{\frac{p^1(1-p^1)}{n}} \qquad (3.24)$$

$$C.L.p^1 = p^1$$

$$U.C.p^1 = \bar{P} + 3\sqrt{\frac{p^1(1-\bar{p})}{n}}$$

The decision criteria remain the same. If the proportion defectives in each sample plotted against sample numbers lie within the control limits, the process is under control otherwise not.

Control limits for the average number of defectives per sample of equal size 'n can also be constructed. If there are K samples drawn at regular intervals from the lot of manufactured products. If p is the proportion of defective, then

The number of defectives in each sample = n.p.

The average number of defective per sample,

$$\bar{np} = \frac{\text{Total number defectives in all}}{k}$$

Now 3 a - control limits can be given as,

$$\text{L.C.L.}_{\bar{p}} = \bar{np} - 3\sqrt{\bar{np}(1-\bar{p})}$$

$$\text{C.L.}_{p^1} = \bar{np} \qquad (3.25)$$

$$\text{U.C.}_{p^1} = \bar{P} + 3\sqrt{\bar{np}(1-\bar{p})}$$

Example 3-3 : The data below gives the number of defectives items in 10 boxes each containing 50 items.

Sample No.:	1	2	3	4	5	6	7	8	9	10
No. of defectives:	4	5	7	3	9	11	8	10	6	2

per box

(i) 3 a control limits for the proportion of defectives can be constructed in the following manner.

Total number of units = 50 x 10 = 500

Total number of defectives = 65

Estimated proportion of defectives, $= \frac{65}{500} = 0.13$

3 σ— control limits from (3.24) are,

$$L.C.L._{\bar{p}} = p^1 - 3\sqrt{\frac{p^1(1-p^1)}{n}} \qquad (3.24)$$

$$C.L._{p^1} = p^1$$

$$U.C._{p^1} = \bar{P} + 3\sqrt{\frac{p^1(1-\bar{p})}{n}}$$

The proportion of defectives in the samples are as given below:

Sample No. :	1	2	3	4	5	6	7	8	9	10
Proportion :	.08	.1	.14	.06	.18	.22	.16	,20	.12	.04

Now we mark the control limits on the graph paper and plot the proportion of defective values against the sample numbers.

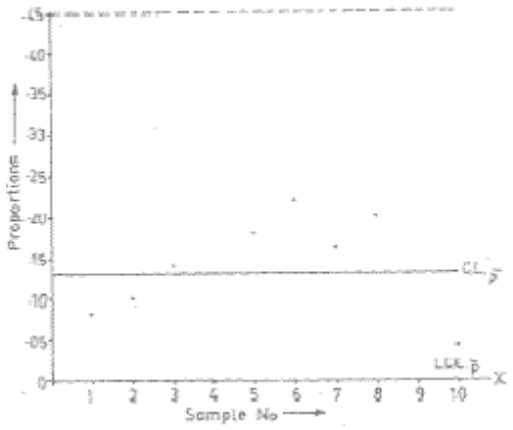


Fig. 23.5 p-chart

From the graph it is apparent that no point lies outside the control limits. Hence, the process is under control.

(ii) Now we establish 3-control limits for number of defectives per sample and test whether the process is under control or not.

From the given data, the average number of defectives per sample are,

$$\bar{np} = \frac{65}{10} 6.5$$

$$\text{and } \bar{p} = \frac{6.5}{50} = 0.13$$

Thus, 3 a - control limits from (3.25) are,

$$\text{L.CL.np} = 6.5 - 3\sqrt{6.5(1-0.13)}$$

$$= 6.5 - 7.13$$

$$= -0.63$$

C.L. np = 6.5

U.C.L.np = 6.5 + 7.13

= 13.63

Now we draw the control limits on the graph and plot the number of defectives against sample numbers as depicted below :

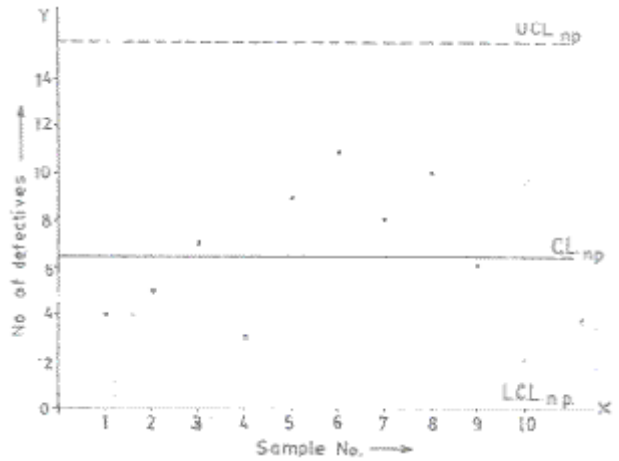


Fig. 23.6 np-chart

The graph reveals that no plotted point lies outside the control band. Hence, we conclude that the process is under control.

Remark:

It should be remembered that if a control limit comes out to be negative it has to be taken equal to zero.

C-chart

If there is any mistake in the item, it is counted as defective. In the counting of defectives, it hardly matters whether an unit has one defect or more defects. But it is also important how many defects per item are there. Hence, we establish the control

limits for the average number of defects per item. The charts used for number of defects are known as C- charts. Since the number of defects per item is a rare event and thus follows Poisson distribution. Hence, the control limits for C-chart are based on Poisson distribution. For instance, number of rivets missing in an aero plane, number of defective seeds in a packet etc.

If the standard value of the nonconformities C is given, then the control limits for C-chart are

$$\begin{aligned} L.C.L.C^1 &= C^1 - 3\sqrt{C^1} \\ C.L.C^1 &= C^1 \\ U.C.L.C^1 &= C^1 + 3\sqrt{C^1} \end{aligned} \qquad (3.26)$$

In case the standard value 'C of nonconformities is hot given, we have to use the estimated value of C which is equal to the average number of defects per unit on the basis of all the samples. If there are K samples and Ci is the number of defects for i sample unit, then the average number of defects per unit is,

$$\bar{C} = \frac{\text{Total no.of defects}}{k} = -\frac{1}{k}\sum_{i-1}^{k} C_i$$

Thus, the 3 a - control limits for the number of defects are:

$$\begin{aligned} L.C.L.C &= \bar{C} - 3\sqrt{C^1} \\ C.L.C &= \bar{C} \\ U.C.L.C &= \bar{C} + 3\sqrt{\bar{C}} \end{aligned} \qquad (3.27)$$

Example 3-4 :

During the inspection of 12 cars, the number of defects were as follows:

Car No.	1	2	3	4	5	6	7	8	9	10	11	12
No. of defects	4	5	7	8	2	3	0	4	6	9	12	10

3a-control limits for the number of defects can be constructed and control chart can be drawn in the following manner.

Since the standard value of C is not known, we estimate it by C where,

$$\bar{C} = \frac{4+5+7+8+2+3+0+4+6+9+12+10}{12}$$

$$= \frac{70}{12}$$

$$= 5.83$$

Hence, the control limits by (3.27) are,

Hence, The control limits by (3.27) are

$$L.C.L._c = 5.83 - 3\sqrt{5.83}$$

$$= 5.83 - 7.24$$

$$= -1.41 = 0$$

$$C.L._c = 5.83$$

$$U.C.L._c = 5.83 + 7.24$$

$$= 13.07$$

Now we demarcate the control limits on the graph paper and plot the number of defects against sample (Car) Nos. as displayed below.

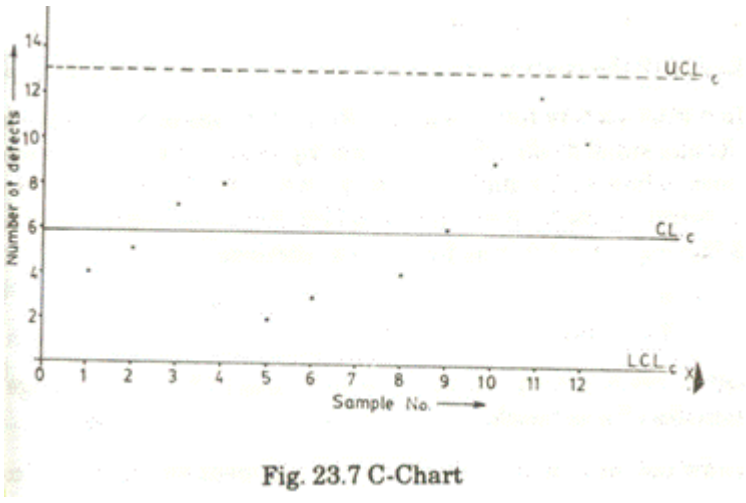


Fig. 23.7 C-Chart

From the chart, it is trivial to draw the conclusion that the process is under control as no plotted point lies above the upper control limits.

Control charts are of great use in industries and big companies. Statistical quality control is of great help to the producer as well as to the buyer.

QUESTIONS

1 Draw the control chart for mean and range for the observations taken on the diameter of lead pellets selected at random as samples of size five and these samples were selected 15 times. The observations were as given below:

Sample No.	Diameter (in mm)				
	I	2	3	4	5
1.	5.2	4.9	5.0	4.6	5.4
2.	4.3	4.7	5.2	4.8	5.3
3.	4.5	4.8	4.9	5.0	5.1
4,	4.9	5.4	5.3	5.2	4.8
5	5.4	5.6	5.6	5.8	5.9
6.	4.8	4.1	5.2	5.3	5.4
7.	4.9	5.1	5.4	5.0	5.2
8.	5.4	5.3	4.9	5.2	5.1
9.	5.0	5.5	5.8	5.3	5.4
10.	5.4	5.5	5.6	5.0	5.0
11.	5.2	4.9	4.7	5.4	5.6
12.	4.8	4.6	4.3	5.2	5.8
13.	5.1	5.4	5.0	5.4	5.9
14.	4.6	4.9	5.2	5.0	4.0
15.	4.9	5.6	4.7	4.3	4.6

Also draw *a* chart for the above data,

2 The number of defects in match boxes of 12 samples drawn from the bundles of match boxes were as follows:

5, 7, 8, 2, 0, 12, 4, 5, 3, 1, 6, 4,

Construct the control limits for the number of defects and draw control chart.

3 The number of defective plugs in the 10 boxes of 20 plugs each were as follows.

2, 3, 4, 0, 1, 5, 2, 3, 6, 4, ...

Establish the control limits for number of defectives and draw p-chart.

4 In a glass factory the task of quality control was done with the help of mean *(X)* and stand deviation a chart 18 samples of 10 items each were chosen and then values of *Ix* and 2s were found to be 595.8 and 8.28 respectively. Determine the 3 limits for standard deviation chart. You may use the following control factors for your calculations:

n	A_1	B_3	B_4
10	1.03	0.28	1.72

5 Given below are the values of sample mean (X) and the range (R) for ten samples of size 5 each.

Draw the mean and range charts and comment on the state of control of the process.

(Given A2 = 0.58, D3 = 0, and D4 = 2.115 for n = 5)

Sample No.	1	2	3	4	5	6	7	8	9	10
X :	43	49	37	44	45	37	51	46	43	47
R :	5	6	5	7	7	4	8	6	4	6

6 In a manufacturing concern producing radio transistors, lots of 250 items are inspected at a time. Considering the number of defectives in 20 lots shown in the table below, draw suitable control chart and write a brief report based on the evidence of the chart.

Lot No	1	2	3	4	5	6	7	8	9	10
No. of defectives :	25	47	23	36	24	34	39	32	35	22
Lot No. :	11	15	13	14	15	16	17	18	19	20
No. of defectives	45	40	32	35	21	40	15	28	2'':'	42

7 A daily sample of 30 items was taken over a period of 14 days in order to establish attributes control limits. If 21 defectives were found, what should be the upper and lower control limits of the proportion of defectives.

8 A drilling machine bores holes with a mean diameter of 0.5230 cm and a standard deviation of 0.0032 cm. calculate the 2-sigma and 3-sigma upper and lower control limits for means of samples 4 and prepare a control charts.

9 Samples of 100 tubes are drawn randomly from the output of a process that produces several thousand units daily. Sample items are inspected for quality and defective tubes are rejected. The results of 15 samples are shown below:

Sample No.	No. of defective tubes	Sample No.	No. of defective tubes
1	8	9	10
2	10	10	13
3	13	11	18
4	9	12	15
5	8	13	12
6	10	14	14
7	14	15	9
8	6		

On the basis of information given above prepare a control chart for fraction defective. What conclusion do you draw from the control chart.

REFERENCES

Agarwal. B.L., *'Basic Statistics',* Wiley Eastern Ltd., New Delhi, 2nd ed., 1991.

Burr, I.W., *'Engineering Statistics and Quality Control',* McGraw-Hill Book Company, New York, 1960.

Cowden, D.J., *'Statistical Methods in Quality Control',* Asia Publishing House, Bombay, 1960.

Duncan, A.J., *'Quality Control and Industrial Statistics',* Richard D.Irwin, Home wood, 1953.

Grant, E.L.,'*Statistical Quality Control',* McGraw-Hill Book Company, NewYork, 1946.

- End Of Chapter -

LESSON – 24

BUSINESS FORECASTING

Introduction

Every businessman wants to be successful in his business. Success of the business can be interpreted as to earn maximum gains and guard against all likelihood losses. This is only possible if he can plan well about his business proceeds. For this most of the middle class businessmen proceed on the basis of their experience and judgement. But this is not all. The big business houses cannot merely depend upon the guess but have to analyse the past data and use is for predictions. Forecasts also forewarn against all the eventualities which are likely to happen so that a businessman can prepare himself. For instance, a slum period is likely to come. So for this, a business man may arrange the finances, godowns etc. and wait for the boom period. Here, we quote some of the writers about their views about forecasting.

Lewis and Fox:

Forecasting is using the knowledge we have at one time to estimate what will happen at some future moment in time.

Neter and Wasserman:

Business forecasting refers to the statistical analysis of the past and current movements in a given time series, so as to obtain clues about the future pattern of the movements.

Leo Barnes:

The aim of forecasting is to establish, as accurately as possible, the probable behaviour of economic activity based on all data available, and to set policies in terms of these probabilities.

H.J. Wheldon:

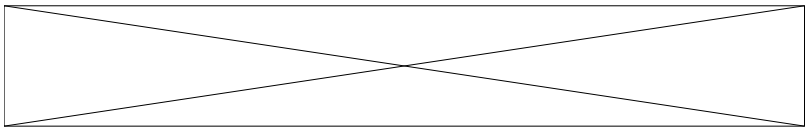
Business forecasting is not so much the estimation of certain figures of sales, production, profits etc. as the analysis of known data, internal and external, in a manner which will enable policy to be determined to meet probable future conditions to the best advantage.

Objectives of the business forecasting

The forecasting is aimed to predict future course of activity on the basis of the analysis of the statistical data available and also the circumstances in which an activity took place. The forecasting is based on minimising the risk of errors in predictions through probability measures.

The forecasts guard a businessman about the future mishappening and also foretell the events which are helpful in his expansion of business. The business activity is not an uniform phenomenon. They are always some periods of booms and depressions. So, through business forecasting one predicts the periods in which a boom is expected and in which a depression is expected as we do in time series analysis also. But in business forecasting one is not confined to statistical analysis only but also takes into consideration the qualitative and circumstantial factors. Now we give some of the forecasting methods barring those which have already been discussed like extrapolation, regression analysis, time series analysis etc. The forecasting techniques are basically classified under three basic categories

i. Naive methods
ii. Barometric methods
iii. Analytical methods



Please use headphones

Each of these methods cover a number of techniques and theories which are given below and discussed one by one.

1) Naive method

i. The economic rhythm theory

2) Barometric methods

i. Specific historical analogy

ii. Lead-lag relationship

iii. Diffusion index

iv. Action-reaction theory

3) Analytical methods

i. The factor listing method
ii. Cross cut analysis theory
iii. Opinion polling
iv. Exponential smoothing
v. Econometric methods

The economic rhythm theory :

Under this theory a businessman analyses the time series data of his own firm for trend, seasonal variation and cycles and forecasts himself. Under this he usually forecasts about his sales, production, dividend etc. The forecast is based on the projection obtained by the analysis of time series data. Such an analysis leads one to decide whether the increase or decrease in demand is cyclical or it is a continued process. The businessman has to adjust his stocks accordingly. If there is a continued increase in demand, a manufacturer can think of expanding his plant or to install a new plant etc. In case of decreasing demand, are may reduce his production or make arrangements to hold the stocks.

We know a time series analysis is not able to exactly forecast the turning point and amplitude of the cycles. Hence, a prominent approach is to predict the cycle of one series by the another series that leads.

The forecasts made by a company under this method are only valid to the company itself and are not applicable to any other business house. Moreover, the forecasts under economic rhythm theory are not very reliable as they do not take into consideration other qualitative factors like government policies, liking of people, etc. *Standard and poors trade and securities service, New York,* have some faith in this theory.

Specific historical analogy :

This technique is based on the principle that history repeats itself. It means a situation occurred in past in economic activity will be repeated in time to come. Hence, it is hoped that a present series will have certain similarities to a past series. Hence, the conclusions drawn from past series are directly applicable to the present series to express the state of economy. The method of historical analogy can be applied to a series in general.

As a matter of fact, it is rare to find a situation in the past which is exactly similar to the present one. Hence, a series in the past is searched out which resemble most. Inferences are drawn to make a forecast giving allowance for dissimilarities in the two situations. This method of forecasting is not very accurate.

Lead-lag relationship method :

This method is also known as *sequence method.* This method is based on the principle that economic activities take place in succession or it can be said that changes occur with a time lag. For instance, in the state of inflation, the exchange rate is adversely affected. The decrease in exchange rate increases the whole sale prices. Which consequently brings rise in retail prices.. When retain prices increase,

people are to be paid higher wages and salaries. So change in one activity of economic phenomenon brings changes in many activities one after the other. Under this theory, the effort is made to determine the time gap between a series and general business cycle. So a turning point in the past leads to forecast what is going to happen to general business activity of connected events.

The lead-lag technique of forecasting takes into account some hypothetical or observed relationship among variables, these relationship are usually established by the inspection of graphs of various series and correlation studies between the series. The famous Harvard index of general business conditions was comprised of three prediction curves (a) *speculation,* the leading series, (b) *business,* the coincident series and (c) *money,* the lagging series. The movement of the three curves is related to one another. The ups and downs of the speculation curve were used to forecast the movement of the business curve. Also speculation and money curve more in opposite directions. The rise in one leads to the fall in the other. Upturn in money curve indicates a recession in business within a few months.

So, in the lead-lag approach three types of indicators are involved namely, lead indicators like exchange rate, money reserves, coincident indicators like employment, industrial production index, total foreign traffic etc. and lag indications like sales, business loans etc.

The main difficulty with lead-lag approach is that it is not easy to interpret the movement of indicators and to establish their relationship. All the more choice of indicators itself is a difficult task. So this method alone is not capable of making reliable forecast but can supplement other methods of forecasting. *National Bureau of Economic Research, Massachusetts* follows lead-lag relationship theory.

Diffusion index:

This method is based on the proposition that all factors affecting a business do not reach their peaks or trough simultaneously. This method does not require to identify which series has a lead and which has a lag. A broad group of series is studied without bothering about any individual series. The diffusion index shows the percentage of the series as expanding or contracting at regular intervals (monthly). So it gives an idea about the general movement of business activity.

All series do not expand or contract simultaneously. One series may be expanding at a point of time while other may be contracting at that time. In such a dilemma, if more than 50 per cent series are expanding, the business is taken to be in the state of booming, otherwise in the state of contracting. The activities of time series are expressed in terms of percentages on monthly basis. National Bureau of Economic Research used 400 time series at a time to determine the diffusion index. Diffusion index are constructed taking any group of variable of business activity like, prices. Profits, prediction, working hours etc. or considering a group of industries.

Action-reaction theory:

This theory is based on Newtons third law of motion that to every action, there is an equal and opposite reaction. So, enough reliance has been placed in action-reaction

theory with regard to business activity as well. Normal condition prevails for longer periods. Any upsurge is bound to follow a recession and a recession will follow an upsurge almost of the same amplitude.

The method faces the difficulty in deciding the line of normal business activity Another problem is to determine the phase through which a company is actually passing at the time of forecast. Even then some forecasting agencies follow this theory for business forecasts. *Business Statistics Organisation* formerly known as Babsons statistical organisation follows action reaction theory.

Factor listing method :

This is a nonmathematical approach for forecasting. In this method, all the factors which are considered to affect a business activity are analysed. Each factor is analyses individually to ascertain whether it is favourable to a business activity or not and then a cumulative picture is drawn by a mental* process. Thus, the forecast is made on the basis of this cumulative picture.

This method is completely flexible with regard to the number of factors tw.

This theory is just opposite to historical analog.

This method is completely flexible with regard to the number of factors their relative importence etc., The method is very easy and handly. But the main drawback of this it is totally subjective. So under this method different persons may give forecasts. Hence, this method is not much trustworthy.

Cross-cut *analysis theory :*

This theory is just opposite to historical analogy approach. Under this theroy it is belived that past cycle cannot be thrusted upon future cycles Hence, past series a cannot be a guide for forecasting. Under this situation, the impact of present policies, demand, technological changes availabiiitv of inputs styles, fads etc. are considered to see their joint impact on economic activity. Also the viewes of economists. Executives and consumers etc. are taken to reach a clear understanding for forecasting. These forecasts are usually short term forecasts.

This method suffers with the lacunae that it is very difficult to see the impact of factors and then to get a cumulative picture. Also people do not see.

Opinion polling

This method of forecasting is entirely based on the opinion of the personnel involved m business like sales officers, production managers executives and also opinion expressed in magazines. The opinion is analysed and the forecast are made. This approach is usually very suitable for short term forecasting.

Exponential smoothing method:

This totally a mathematical approach of forecasting .The trend by moving average method had been discussed undertime series analysis where equal weights are assigned to all item. But under this method, the weights are assigned which are in geometric progression . More weightage is given to recent observation and less to distant observation.

Suppose the weights assigned to n observations are

1, (1-w) , $(1-w)^2$.....$(1-w)^{n-1}$ for $0 < w < 1$

The weighted average till the current period it is

1, (1-w) , $(1-w)^2$.....$(1-w)^{n-1}$ for $0 < w < 1$

The weighted average till the current period it is

$$X_t = \frac{1X_t + (1-w)X_{t-1} + (1-w)^2 X_{t-2} + \ldots + (1-w)^{n-1} X_{t-n+1}}{1 + (1-w) + (1-w)^2 + \ldots + (1-w)^{n-1}} \qquad (4.1)$$

Similarly for the period (t+1), the weighted average is,

$$X_{t+1} = \frac{1X_t + 1 + (1-w)X_t + (1-w)^2 X_{t-1} + \ldots + (1-w)^{n-1} X_{t-n+2}}{1 + (1-w) + (1-w)^2 + \ldots + (1-w)^{n-1}} \qquad (4.2)$$

Taking n large and neglecting higher power of w and (1-w) and doing algebraic manipulation, the relation between X_t+1 with the help of (4.1) and (4.2) can be established as

$$X_{t+1} = w X_{t+1} + (1-w) X_t \qquad (4.3)$$

In (4.3), X_{t+i} is known as the smoothed value at the time (t+1) which is a to make forecast the smoothed value Now to make a forecast, the smoothed values are used to find out a change in each period. The constant w is called the smoothing coefficient and (1-w) /w the trend factor.

Since in the above process only one constant w is used, it is known as the *single parameter exponential smoothing.* The forecast for the first period is taken from some old forecast or is assumed by him.

Trend adjusted exponential smoothing

A situation is also faced where trend and forecast move in opposite direction. To minimize this effect,the calculation under trend adjusted exponential smoothing are made. Here the forecasts are adjusted according to the trend.

Now we rewrite the equation (4.3)

$X_{t+1}=w(X_{t+1}X_t)+X_t$

$X = w(X_t-X_{t-1})+X_{t-1}$ (4.4)

The quantity (X_t-X_{t-1}) is called the error. Thus from (4.4), the forecast at time t is the sum of the proceeding forecast and w times the error. The trend coefficient required for doing the forecast is obtained by the formula, Trend coefficient,O_t=wx change in smoothed value +(l-w)x preceding trend coefficient.

Thus the forecast,

Ft = Smoothed value + Trend factor x Trend coefficient. (4.6)

The error in forecast,

Et = Xt-Ft (4.7)

Choice of the constant w :

W is chosen arbitrarily. There is no rule which governs to choose a value of w. Anyhow guidance may be provided. If the fluctuations in economic factors like, sales, production, demand, profits etc. are of random nature, a small value of w is preferred. Also if the actual value turns down and forecast do not turn down, a large value of w is usually chosen and vice-versa.

Example 4-1 :

Monthly pattern of off-take of edible oils by the public distribution system during the year 1989-90 were as follows:

Months	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Edible oils : ('000 Tonnes)	24.2	24.4	23.0	28.8	34.9	37.7	38.3	42.9	26.0

Month	Jan.	Feb.	Mar
Edible oils : ('000 Tonnes)	27.0	28.5	35.0

The smoothed values and trend adjusted forecasts for the above time series can uc calculated in the following manner. Let us take w = 0.4. The values are calculated step by step and entered in the table given below:

$$\text{Since W} = 0.4, \text{the trend factor } \frac{1-w}{w} = \frac{.6}{.4} = \frac{3}{2}$$

Also suppose that the first smoothed value X = 25.0 smoothed forecasts from May to Mar by the formula (4.4) are,

May, X_2 = .4(24.4 - 25.000) 4-25 = 24.76

June, X_3 = .4(23.0 - 24.760) + 24.76 = 24.056

July, X_4 = .4(28.8- 24.056) + 24.056 = 25.954

Aug., X_5 = .4(34.9 - 25.954) + 25.954 = 29.532

Sep., X_6 = .4(37.7-23.532) + 29.532 = 32.799

Oct, X_7 = .4(38.3 - 32.799) + 32.799 = 34.999

Nov.,X_8 = .4(42.9-34.999) -f 34.999 = 38.159

Dec.,X9 =.4(26.0-38.159) +38.159 = 33.295

Jan.,X1o =.4(27.0-33.295) +33.295 = 30.777

Feb.,X11 = .4(28.5-30.777) + 30.777 = 29.866

Mar.,2T12 = .4(35.0-29.866) + 29.866 = 31.920

Month	Observed value	Smoothed value	Change in smoothed value	Trend co-efficient	Forecast	Error
	Xt	**Xt**	**Xt+1 - Xt**	**θ_t**	**Ft**	**Et**
Apr.	24.2	25.0		_	_	_
May	24.4	24.76	-0.24	-0.096	24.616	-0.216
Jun.	23.0	24.056	- 0.704	-0.339	23.548	-548
Jul	28.8	25.954	1.898	0.559	26.792	2.008
Aug.	34.S	29.532	3.578	1.767	32.182	2.718
Sept.	'37.7	32.799	3.267	2.367	36.350	1.350
Oct.	38.3	34.999	2.200	2.300	38.449	-0.149
Nov.	42.9	38.159	3.160	2.644	• 42.125	0.780
Dec.	26.0	32.295	-5.864	- 0.759	31.156	-5.156
Jan.	27.0	30.777	-1.518	-1.063	29.182	-2.182
Feb.	28.5	29.866	-0.911	; -1.002	28.363	0.137
Mar.	35.0	31.920	2.054	0.220	32.25	2.750

Change in smoothed values Xt+i - Xt is calculated by subtracting the smoothed value of the preceding period from its next period.

For Apr. not available.

For May, change in smoothed value = 24.76-25.0 =-0.24

For June, change in smoothed value = 24.056 - 24.76 = -0.704

. .

. .

. .

. .

For Mar., change in smoothed value = 31.92 - 29.866 = 2.054

Trend coefficient from (4.5) are calculated below:

For Apr., not available,

For May, 02 = *A* x (-0.24) + (1-.4) x 0-0.096

For June 03 = .4x (-0.704) + (l-.4)x -0.096) = - 0.339

For July 04 = .4 x 1.898 +(1- .4) (- 0.339) = 0.556

. .

. .

. .

. .

For forecast F_t, the trend factor $\frac{1-.4}{.4}1.5$ So,

For Apr., Fi is not available

For May F_2 = 24.76 + 1.5 x (-0.096) = 24.616

For June, Fs = 24.056 + 1.5 x (-0.339) - 23.548

. .

. .

. .

Now, the error Et is calculated as,

For May, E2 = 24.4 - 24.616 = - 0.216

For June, E3 = 23.0 - 23.548 = - 0.548

For July, E_4 = 28.8 - 26.792 = 2.008

. .

. .

. .

Note:

Exponential smoothing can be applied with more than one trend coefficient also. But the details are kept out of this chapter.

Econometric method :

The manner is which an economic system behaves depends on a number of variables that influence it. The inter- relationship between the variables is expressed by a set of equations. These variables are of two types in nature namely (i) Endogenous variables, (ii) Exogenous variables. Endogenous variables are those which belong to the economic system itself like production, sales, prices, employment, wages etc. Again, exogenous variables are those which affect the economic system but do not belong to it like politics, fads, styles etc. But econometric methods deals with the quantitative variables only which influence and economic phenomenon. On the basis of economic analysis one can forecast about economic changes with certain probability Level.

Econometric analysis fundamentally concerns with the model building. Building a model means estimation of parameters involved in the model through the time series data as we do in a way in regression and time series analysis. There are hundreds of econometric models applicable in various areas of economic activity. Here, we discuss only one model just to explain the idea lying behind econometric methods.

Now we consider a model for Gross National Product (GNP) at a time period t.

Y_t = Ct + It + Gt ' (4.8)

Where,

Yt = GNP at time t

Ct = Consumption for time period t

It = Gross investment at time t

Gt = Government expenditure in time t.

In model (4.8), each factor on the right hand side is a function of other variables. We know, the consumption Ct is a function of the increment in consumption, irrespective of the initial, value of GNP. The incremental value is known as *marginaL propensity* and is denoted by . Also there will be some consumption even if the G NP is zero. Let it be denoted by . Thus, the model for Ct is,

C_t = + a+b Y_{t-1}. (4.9)

The gross investment is equal to the total investment made by the private sector and autonomous bodies. The induced investment in any period t is proportion to the difference in consumption in period t and its presiding equivalent period. it be denoted i. Also the amount invested for replacement and adoption of new technology comes under the category of autonomous investment. Let this amount be denoted by K.Hence, the model for It is,

It = k + i(C_t-C_t-i)(4.10)

Government expenditure is an exogenous variable and is fixed. Let this amount be

Therefore,

Gt = G_0 (4.11)

Now the model (4.8) after substituting the values of C_t, I_t and G_t from (4.9), (4.10) become,

Again,

Ct-1=α+β Yt+1

$$Y = \alpha + \beta\ Y_{t-1} + k + i\ (G_t - G_{t-1}) + G_o$$

$$Y = \alpha + \beta\ (1+i)\ Y_{t-1} i\ \beta Y_{t-2} + k + G_o \qquad (4.13)$$

Reduce model (4.13) involves four parameters , , K and I . These parameters can be estimated on the basis of time series data in the usual manner , Let these

$$Y = a + b\ (1+i)\ Y_{t-1} i\ b Y_{t-2} + k + G_o \qquad (4.14)$$

Substituting the values of Yt, Y_{t-1}, Y_{t-2} and G_0 for any time period t in the equation (4.14) we obtain the values of Y_t for the forecast period t.

Econometric method are good and reliable but not exact due to sampling errors. All the more , a forecast will be good only when the econometric model is suitable for the data. Also , the forecast are not the same as they are likely to be in times to come. Anyhow still they provide some good guidelines to planning.

Forecasting in India has not been taken up professionally and is still in an infant stage . The business houses are realizing its importance and establishing cells in their own companies.

QUESTIONS

1. Throw light on the importance of business forecasting.

2. Write a short note on business forecasting.

3. How does an econometric model differ from a regression model?

4. Is it possible to develop an infallible system of forecasting? Justify your answer.

5. Name three forecasting agencies.

6. Describe the method of exponential smoothing for forecasting.

7. Give lead-lag method of forecasting.

8. Differentiate between long term and short term forecasting. Name the method, wh.ch are suitable for long term and short term forecasting

9. Examine critically the time lag and the action-reaction theory of business forecasting. Which of these, in your opinion is better and why?

10 Per capita consumption of tea (in gms from 1976-77 to 1986- 87 is given in the table below:

Years :	1976-77	1977-78	1978-79	1979-80	1980-81	1951-82
	455	469	479	498	518	487
	1982-83	1983-84	1984-85	1985-86	1986-87	
	464	367	399	422	420	

Find the forecast values by the method of exponential smoothing taking the initial forecast as 450 and w = 0.5.

ll Monthly pattern of off-take of wheat *by* public distribution system is as follows :

Months	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.
Wheat: ('000 Tonnes	544 les)	525	604	655	614	643	626	619

('000 Tonnes)

Mqiith :	Dec.	Jan.	Feb.	Mar.
Wheat:	724	684	662	768

Find the forecast values by exponential smoothing method taking w = 0.6 and initial forecast as 525.

REFERENCES

Agarwal, B.L., *'Basic Statistics',* Wiley Eastern Ltd., New Delhi, 2nd ed., 1991.

Chou, Y.L., *'Applied Business and Economic Statistics',* Holt, Rinehart and Whiston, New York, 1963.

Firth, M., *'Forecasting Methods in Business and Management',* Edward Arnold, London, 1977

Wheel, W, and Makridakis, . *'Forecasting:* Methods and Application', John Wiley, London, 1978.

- End Of Chapter -