F.Y.B.SC (I.T)

SEMESTER - IV

PAPER - IV

QUANTITATIVE TECHNIQUES

Dr. Suhas Pednekar
Vice Chancellor
University of Mumbai, Mumbai

Prof. Ravindra D. Kulkarni
Pro Vice-Chancellor,
University of Mumbai

Prof. Prakash Mahanwar
Director,
IDOL, University of Mumbai

Programme Co-ordinator : **Shri Mandar Bhanushe**
Head, Faculty of Science and Technology IDOL
Univeristy of Mumbai – 400098

Course Co-ordinator : **Gouri S.Sawant**
Assistant Professor B.Sc.I.T, IDOL
University of Mumbai- 400098

Course Writers : **Dr. Manish Pithadia**
Assistant Professor,
Viva College of Arts, Science and Commerce.

April 2021, Print - 1

Published by : **Director,**
Institute of Distance and Open Learning,
University of Mumbai,
Vidyanagari,Mumbai - 400 098.

DTP Composed by : **7SKILLS**
Dombivli West, Thane - 421202

Printed by :

CONTENTS

F.Y. B.Sc. (IT)
Semester - IV

Quantitative Techniques

SYLLABUS

Unit 1 **Errors, Solutions of Algebraic and Transcendental Equations** using - Bisection Method, the Method of False Position, Newton-Raphson Method.

Interpolation: Interpolation: - Forward Difference, Backward Difference, Newton's Forward Difference Interpolation, Newton's Backward Difference Interpolation, Lagrange's Interpolation

Unit 2 **Solution of simultaneous algebraic equations (linear) using iterative methods:** Gauss-Jordan Method, Gauss-Seidel Method. **Numerical Integration:** Trapezoidal Rule, Simpson's 1/3 rd and 3/8 th rules. **Numerical solution of 1st and 2nd order differential equations: -** Taylor **series,** Euler's Method, Modified Euler's Method, Runge-Kutta Method for 1st and 2nd Order Differential Equations.

Unit 3 **Random variables:** Discrete and Continuous random variables, Probability density function, Probability distribution of random variables, Expected value, Variance.

Moments and moment generating functions: Relation between Raw moments and Central moments.

Distributions: Binomial, Poisson, Normal, exponential, uniform distributions for detailed study, Central Limit theorem (statement only) and problems based on this theorem.

Unit 4 **Fitting of curves:** Least square method, Fitting the straight line and parabolic curve, Correlation, Covariance, Karl Pearson's coefficient and Spearman' s Rank, correlation coefficient, Regression coefficients and lines of regression.

Unit 5 **Sampling distribution:** Test of Hypothesis, Level of Significance, Critical Region, One Tailed and Two Tailed Test, Interval Estimation of Population Parameters, Test of Significance for large Samples and small Samples, Student's 't' Distribution and its properties.

Unit 6 **Chi-Square Distribution and its properties,** Test of the Goodness of Fit and Independence of Attributes, Contingency Table, Yates Correction

Mathematical Programming: Linear optimization problem, Formulation and Graphical solution, Basic solution and Feasible solution, Primal Simplex Method

Books:

Introductory Methods of Numerical Methods, Vol-2, S.S.Shastri, PHI
Fundamentals of Mathematical Statistics, S.C.Gupta, V.K.Kapoor

Reference:

Elements of Applied Mathematics, Volume 1 and 2, P.N.Wartikar and J.N.Wartikar, A. V. Griha, Pune

Engineering Mathematics, Vol-2, S.S.Shastri, PHI

Applied Numerical Methods for Engineers using SC/LAB and C, Robert J.Schilling and Sandra

L.Harris, ", Thomson Brooks/Cole

UNIT 1

1

SOLUTIONS OF ALGEBRAIC AND TRANSCENDENTAL EQUATIONS

Unit Structure

1.0 Objectives

This chapter will enable the learner to:

- understand the concepts of simple root, multiple roots, algebraic equations, transcendental equations, direct methods, iterative methods.

- find roots of an equation using Bisection method, Newton-Raphson method, Regula-Falsi method, Secant method.
- understand the geometrical interpretation of these methods and derive the rate of convergence.

1.1 Introduction

The solution of an equation of the form $f(x) = 0$ is the set of all values which when substituted for unknowns, make an equation true. Finding roots of an equation is a problem of great importance in the fields of mathematics and engineering. In this chapter we see different methods to solve a given equation.

1.1.1 Simple and Multiple Roots

Definition 1.1.1.1 (Root of an Equation). A number ζ is said to be a root or a zero of an equation $f(x) = 0$ if $f(\zeta) = 0$.

Definition 1.1.1.2 (Simple Root). A number ζ is said to be a simple root of an equation $f(x) = 0$ if $f(\zeta) = 0$ and $f'(\zeta) \neq 0$. In this case we can write $f(x)$ as

$$f(x) = (x - \zeta)g(x), \text{ where } g(\zeta) \neq 0.$$

Definition 1.1.1.3 (Multiple Root). A number ζ is said to be a multiple root of multiplicity m of an equation $f(x) = 0$ if $f(\zeta) = 0, f'(\zeta) = 0, \ldots, f^{(m-1)}(\zeta) = 0$ and $f^{(m)}(\zeta) \neq 0$. In this case we can write $f(x)$ as

$$f(x) = (x - \zeta)^m g(x), \text{ where } g(\zeta) \neq 0.$$

1.1.2 Algebraic and Transcendental Equations

Definition 1.1.2.1 (Algebraic Equation). A polynomial equation of the form $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1} + a_nx^n = 0$, $a_n \neq 0$ where $a_i \in \mathbb{C}$ for $0 \leq i \leq n$ is called an algebraic equation of degree n.

Definition 1.1.2.2 (Transcendental Equation). An equation which contains exponential functions, logarithmic functions, trigonometric functions etc. is called a transcendental function.

1.1.3 Direct Methods and Iterative Methods

Definition 1.1.2.1 (Direct Methods). A method which gives an exact root in a finite number of steps is called direct method.

Definition 1.1.2.2 (Iterative Methods). A method based on successive approximations, that is starting with one or more initial approximations to the root, to obtain a sequence of approximations or iterates which converge to the root is called an iterative method.

1.1.4 Intermediate Value Theorem

Iterative methods are based on successive approximations to the root starting with one or more initial approximations. Choosing an initial approximation for an iterative method plays an important role in solving the given equation in a smaller number of iterates. Initial approximation to the root can be taken from the physical considerations of the given problem or by graphical methods.

As an example of finding initial approximation using physical considerations of the given problem, consider $f(x) = x^3 - 28$. Then to find the root of $f(x) = 0$, one of the best initial approximation is $x = 3$ as cube of $x = 3$ is close to the given value 28

To find initial approximation graphically, consider an example of

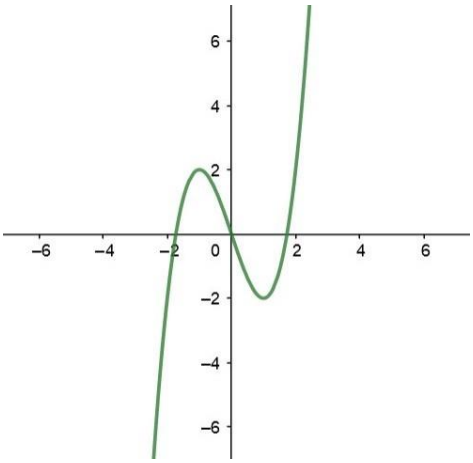


Figure 1: $f(x) = x^3 - 3x$

$f(x) = x^3 - 3x$. The value of x at which the graph of $f(x)$ intersects x- axis gives the root of $f(x) = 0$. From Figure 1, it is clear that the roots of $f(x) = 0$ lies close to 2 and −2. Hence the best initial approximations will be 2 and −2.

Intermediate Value Theorem is another commonly used method to obtain the initial approximations to the root of a given equation.

Theorem 1.1.4.1 (Intermediate Value Theorem). If $f(x)$ is a continuous function on some interval $[a,b]$ and $f(a)f(b) < 0$ then the equation $f(x) = 0$ has at least one real root or an odd number of real roots in the interval (a,b).

1.1.5 Rate of Convergence

An iterative method is said to have a rate of convergence α, if α is the largest positive real number for which there exists a finite constant $C \neq 0$ such that

$$|\epsilon_{k+1}| \leq C|\epsilon_k|^{\alpha}$$

where $\epsilon_i = x_i - \zeta$ is the error in the i^{th} iterate for $i \in \mathbb{N} \cup \{0\}$. The constant C is called the asymptotic error constant.

1.2 Bisection Method

Bisection method is based on the repeated application of intermediate value theorem. Let $I_0 = (a_0, b_0)$ contain the root of the equation $f(x) = 0$. We find $m_1 = \frac{a_0+b_0}{2}$ by bisecting the interval I_0. Let I_1 be the interval (a_0, m_1), if $f(a_0)f(m_1) < 0$ or the interval (m_1, b_0), if $f(m_1)f(b_0) < 0$. Thus interval I_1 also contains the root of $f(x) = 0$. We bisect the interval I_1 and take I_2 as the subinterval at whose end points the function $f(x)$ takes the values of opposite signs and hence I_2 also contains the root.

Repeating the process of bisecting intervals, we get a sequence of nested subintervals $I_0 \supset I_1 \supset I_2 \supset \cdots$ such that each subinterval contains the root. The midpoint of the last subinterval is taken as the desired approximate root.

Example 1.2.1. Find the smallest positive root of $f(x) = x^3 - 5x + 1 = 0$ by performing three iterations of Bisection Method.

Solution: Here, $f(0) = 1$ and $f(1) = -3$. That is $f(0)f(1) < 0$ and hence the smallest positive root lies in the interval $(0,1)$. Taking $a_0 = 0$ and $b_0 = 1$, we get

(First Iteration)

$$m_1 = \frac{a_0 + b_0}{2} = \frac{0+1}{2} = 0.5$$

Since, $f(m_1) = -1.375$ and $f(a_0)f(m_1) < 0$, the root lies in the interval $(0, 0.5)$.

Taking $a_1 = 0$ and $b_1 = 0.5$, we get

(Second Iteration)

$$m_2 = \frac{a_1 + b_1}{2} = \frac{0+0.5}{2} = 0.25$$

Since, $f(m_2) = -0.234375$ and $f(a_1)f(m_2) < 0$, the root lies in the interval $(0, 0.25)$.

Taking $a_2 = 0$ and $b_2 = 0.25$, we get

(Third Iteration)

$$m_3 = \frac{a_2 + b_2}{2} = \frac{0 + 0.25}{2} = 0.125$$

Since, $f(m_3) = 0.37695$ and $f(m_3)f(b_2) < 0$, the approximate root lies in the interval $(0.125, 0.25)$.

Since we have to perform three iterations, we take the approximate root as midpoint of the interval obtained in the third iteration, that is $(0.125, 0.25)$. Hence the approximate root is 0.1875.

1.3 Newton-Raphson Method

Newton-Raphson method is based on approximating the given equation $f(x) = 0$ with a first degree equation in x. Thus, we write

$$f(x) = a_0 x + a_1 = 0$$

whose root is given by $x = -\frac{a_1}{a_0}$ such that the parameters a_1 and a_0 are to be determined. Let x_k be the k^{th} approximation to the root. Then

$$f(x_k) = a_0 x_k + a_1 \quad (1.3.1)$$

and

$$f'(x_k) = a_0 \quad (1.3.2)$$

Substituting the value of a_0 in 1.3.1 we get $a_1 = f(x_k) - f'(x_k)x_k$. Hence,

$$x = -\frac{a_1}{a_0} = -\frac{f(x_k) - f'(x_k)x_k}{f'(x_k)}.$$

Representing the approximate value of x by x_{k+1} we get

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, \cdots (1.3.3)$$

This method is called the Newton-Raphson method to find the roots of $f(x) = 0$.

Alternative:

Let x_k be the k^{th} approximation to the root of the equation $f(x) = 0$ and h be an increment such that $x_k + h$ is an exact root. Then

$$f(x_k + h) = 0.$$

Using Taylor series expansion on $f(x_k + h)$ we get,

$$0 = f(x_k + h) = f(x_k) + hf'(x_k) + \frac{1}{2!}h^2 f''(x_k) + \cdots$$

Neglecting the second and higher powers of h we get,

$$f(x_k) + h f'(x_k) = 0.$$

Thus,

$$h = -\frac{f(x_k)}{f'(x_k)}$$

We put $x_{k+1} = x_k + h$ and obtain the iteration method as

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, \cdots$$

Newton-Raphson method requires two function evaluations f_k and f_k' for each iteration.

1.3.1 Geometrical Interpretation

We approximate $f(x)$ by a line taken as a tangent to the curve at x_k which gives the next approximation x_{k+1} as the x-intercept as in Figure 2.

Example 1.3.1. Find the approximate root correct upto two decimal places for $f(x) = x^4 - x - 10$ using Newton-Raphson Method with initial approximation $x_0 = 1$.

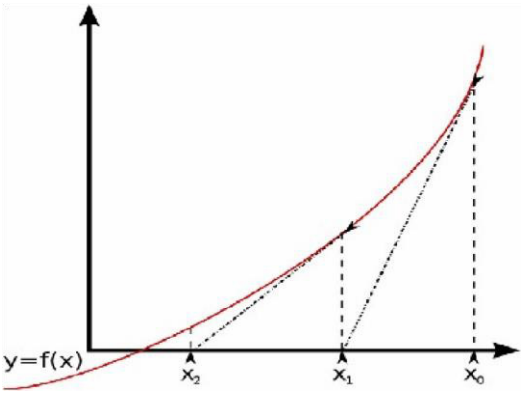


Figure 2: Newton-Raphson Method

Solution: Here $f(x) = x^4 - x - 10 = 0$ implies $f'(x) = 4x^3 - 1$. For $x_0 = 1$, $f(x_0) = -10$ and $f'(x_0) = 3$. By Newton-Raphson iteration Formula,

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

Thus,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 4.3333$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 3.2908.$$

Similarly, $x_3 = 2.5562$, $x_4 = 2.0982$, $x_5 = 1.8956$, $x_6 = 1.8568$, $x_7 = 1.8555$.

1.3.2 Rate of Convergence

Let $f(x)$ be a continuous function. Then the Newton-Raphson method to find the approximate root of $f(x) = 0$ is given by equation 1.3.3.

Let ζ be the exact root of $f(x)$. Then we write x_{k+1}, x_k in terms of ζ as

$x_{k+1} = \zeta + \epsilon_{k+1}$ and $x_k = \zeta + \epsilon_k$, where I is the error in the i^{th} iteration.

Let

$$f_k = f(x_k) = f(\zeta + \epsilon_k).$$

Thus, equation 1.3.3 becomes

$$\zeta + \epsilon_{k+1} = \zeta + \epsilon_k - \frac{f(\zeta + \epsilon_k)}{f'(\zeta + \epsilon_k)}.$$

That is

$$\epsilon_{k+1} = \epsilon_k - \frac{f(\zeta + \epsilon_k)}{f'(\zeta + \epsilon_k)}. \qquad (1.3.4)$$

Using Taylor series expansion on $f(\zeta + \epsilon_k)$ and $f'(\zeta + \epsilon_k)$, we get

$$f(\zeta + \epsilon_k) = f(\zeta) + \epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots$$

and

$$f'(\zeta + \epsilon_k) = f'(\zeta) + \epsilon_k f''(\zeta) + \frac{1}{2!}\epsilon_k^2 f'''(\zeta) + \cdots$$

Thus, equation 1.3.4 becomes

$$\begin{aligned}\epsilon_{k+1} &= \epsilon_k - \frac{\left[f(\zeta) + \epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots\right]}{\left[f'(\zeta) + \epsilon_k f''(\zeta) + \frac{1}{2!}\epsilon_k^2 f'''(\zeta) + \cdots\right]} \\ &= \epsilon_k - \left[\epsilon_k + \frac{1}{2!}\frac{f''(\zeta)}{f'(\zeta)}\epsilon_k^2 + \cdots\right]\left[1 + \frac{f''(\zeta)}{f'(\zeta)}\epsilon_k + \frac{1}{2!}\frac{f'''(\zeta)}{f'(\zeta)}\epsilon_k^2 + \cdots\right]^{-1} \\ &= \epsilon_k - \left[\epsilon_k + \frac{1}{2!}\frac{f''(\zeta)}{f'(\zeta)}\epsilon_k^2 + \cdots\right]\left[1 - \frac{f''(\zeta)}{f'(\zeta)}\epsilon_k - \frac{1}{2!}\frac{f'''(\zeta)}{f'(\zeta)}\epsilon_k^2 + \cdots\right]\end{aligned}$$

Neglecting the third and higher powers of ϵ_k, we get

$$\begin{aligned}\epsilon_{k+1} &= \frac{1}{2}\frac{f''(\zeta)}{f'(\zeta)}\epsilon_k^2 \\ &= C\epsilon_k^2\end{aligned}$$

where $C = \frac{1}{2}\frac{f''(\zeta)}{f'(\zeta)}$. Thus, Newton-Raphson method has a second order rate of convergence.

1.4 Regula-Falsi Method

Given a continuous function $f(x)$, we approximate it by a first-degree equation of the form $a_0x + a_1$, such that $a_0 \neq 0$ in the neighbourhood of the root. Thus, we write

$$f(x) = a_0x + a_1 = 0 \tag{1.4.1}$$

Then

$$f(x_k) = a_0x_k + a_1 \text{ and} \tag{1.4.2}$$

$$f(x_{k-1}) = a_0x_{k-1} + a_1 \tag{1.4.3}$$

On solving equation 1.4.2 and 1.4.3, we get

$$a_0 = \frac{f_k - f_{k-1}}{x_k - x_{k-1}}$$

and

$$a_1 = \frac{x_kf_{k-1} - x_{k-1}f_k}{x_k - x_{k-1}}$$

writing x as x_{k+1} we get

$$x_{k+1} = -\frac{\left(\frac{x_k f_{k-1} - x_{k-1} f_k}{x_k - x_{k-1}}\right)}{\left(\frac{f_k - f_{k-1}}{x_k - x_{k-1}}\right)}$$

That is

$$x_{k+1} = \frac{f_k x_{k-1} - f_{k-1} x_k}{f_k - f_{k-1}} \tag{1.4.4}$$

which can be expressed as

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f_k - f_{k-1}} f_k, \qquad k = 1, 2, \cdots \tag{1.4.5}$$

Here, we take the approximations x_k and x_{k-1} such that $f(x_k)f(x_{k-1}) < 0$. This method is known as Regula Falsi Method or False Position Method. This method requires only one function evaluation per iteration.

Example 1.4.1. Perform four iterations of Regula Falsi method for $f(x) = x^3 - 5x + 1$ such that the root lies in the interval $(0,1)$.

Solution: Since the root lies in the interval $(0,1)$, we take $x_0 = 0$ and $x_1 = 1$. Then $f(x_0) = f(0) = 1$ and $f(x_1) = f(1) = -3$. By Regula Falsi Method,

$$x_2 = x_1 - \frac{x_1 - x_0}{f_1 - f_0} f_1 = 0.25$$

and $f(0.25) = -0.234375$

As $f(x_0)f(x_2) < 0$ and $f(x_1)f(x_2) > 0$, by Intermediate Value property, the root lies in the interval (x_0, x_2). Hence

$$x_3 = x_2 - \frac{x_2 - x_0}{f_2 - f_0} f_2 = 0.202532$$

and $f(x_3) = -0.004352297$.

Similarly, we get $x_4 = 0.201654$ and $x_5 = 0.20164$. Hence, we get the approximate root as 0.20164.

1.4.2 Rate of Convergence

Let $f(x)$ be a continuous function. Then the Regula Falsi method to find the approximate root of $f(x) = 0$ is given by equation 1.4.5. Let ζ be the exact root of $f(x)$. Then we write x_{k+1}, x_k, x_{k-1} in terms of ζ as

$$x_{k+1} = \zeta + \epsilon_{k+1}, x_k = \zeta + \epsilon_k, x_{k-1} = \zeta + \epsilon_{k-1}$$

where ϵ_i is the error in the i^{th} iteration.

Hence, $f_k = f(x_k) = f(\zeta+\epsilon_k)$, $f_{k-1} = f(x_{k-1}) = f(\zeta+\epsilon_{k-1})$.

Thus, equation 1.4.5 becomes

$$\zeta + \epsilon_{k+1} = \zeta + \epsilon_k - \frac{\zeta + \epsilon_k - (\zeta + \epsilon_{k-1})}{f(\zeta + \epsilon_k) - f(\zeta + \epsilon_{k-1})} f(\zeta + \epsilon_k)$$

$$\epsilon_{k+1} = \epsilon_k - \frac{\zeta + \epsilon_k - (\zeta + \epsilon_{k-1})}{f(\zeta + \epsilon_k) - f(\zeta + \epsilon_{k-1})} f(\zeta + \epsilon_k)$$

Applying Taylor expansion on $f(\zeta + \epsilon_k)$ and $f(\zeta + \epsilon_{k-1})$ we get

$$\epsilon_{k+1} = \epsilon_k - \frac{(\epsilon_k - \epsilon_{k-1})\left[f(\zeta) + \epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots\right]}{\left[f(\zeta) + \epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots\right] - \left[f(\zeta) + \epsilon_{k-1} f'(\zeta) + \frac{1}{2!}\epsilon_{k-1}^2 f''(\zeta)\right.}$$

Since ζ is an exact root of $f(x) = 0$ we get

$$\begin{aligned}
\epsilon_{k+1} &= \epsilon_k - \frac{(\epsilon_k - \epsilon_{k-1})\left[\epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots\right]}{\left[\epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots\right] - \left[\epsilon_{k-1} f'(\zeta) + \frac{1}{2!}\epsilon_{k-1}^2 f''(\zeta) + \cdots\right]} \\
&= \epsilon_k - \frac{(\epsilon_k - \epsilon_{k-1})\left[\epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots\right]}{(\epsilon_k - \epsilon_{k-1}) f'(\zeta) + \frac{1}{2!}(\epsilon_k^2 - \epsilon_{k-1}^2) f''(\zeta) + \cdots} \\
&= \epsilon_k - \left[\frac{\epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots}{f'(\zeta) + \frac{1}{2!}(\epsilon_k + \epsilon_{k-1}) f''(\zeta) + \cdots}\right] \\
&= \epsilon_k - \left[\frac{\epsilon_k + \frac{1}{2!}\frac{\epsilon_k^2 f''(\zeta)}{f'(\zeta)} + \cdots}{1 + \frac{(\epsilon_k + \epsilon_{k-1}) f''(\zeta)}{2! f'(\zeta)} + \cdots}\right] \\
&= \epsilon_k - \left[\epsilon_k + \frac{1}{2!}\frac{\epsilon_k^2 f''(\zeta)}{f'(\zeta)} + \cdots\right]\left[1 + \frac{(\epsilon_k + \epsilon_{k-1}) f''(\zeta)}{2! f'(\zeta)} + \cdots\right]^{-1} \\
&= \epsilon_k - \left[\epsilon_k + \frac{1}{2!}\frac{\epsilon_k^2 f''(\zeta)}{f'(\zeta)} + \cdots\right]\left[1 - \frac{(\epsilon_k + \epsilon_{k-1}) f''(\zeta)}{2! f'(\zeta)} + \cdots\right]
\end{aligned}$$

Neglecting the higher powers of ϵ_k and ϵ_{k-1}, we get

$$\epsilon_{k+1} = \frac{\epsilon_k \epsilon_{k-1} f''(\zeta)}{2! f'(\zeta)}$$

Hence

$$\epsilon_{k+1} = C \epsilon_k \epsilon_{k-1} \tag{1.4.6}$$

where $C = \frac{f''(\zeta)}{2! f'(\zeta)}$.

Since in Regula Falsi method, one of the x_0 or x_1 is fixed, equation 1.4.6 becomes

$\epsilon_{k+1} = C\epsilon_0\epsilon_k = C'\epsilon_k$, if x_0 is fixed and $\epsilon_{k+1} = C\epsilon_1\epsilon_k = C'\epsilon_k$, if x_1 is fixed.

In both the cases, the rate of convergence is 1. Thus, Regula Falsi method has a linear rate of convergence.

1.5 Secant Method

Given a continuous function $f(x)$, we approximate it by a first-degree equation of the form $a_0x + a_1$, such that $a_0 \neq 0$ in the neighbourhood of the root. Thus, we write

$$f(x) = a_0x + a_1 = 0 \qquad (4.5.1)$$

Then

$$f(x_k) = a_0x_k + a_1 \quad (1.5.2)$$

and

$$f(x_{k-1}) = a_0x_{k-1} + a_1 \qquad (1.5.3)$$

On solving equation 1.5.2 and 1.5.3, we get

$$a_0 = \frac{f_k - f_{k-1}}{x_k - x_{k-1}}$$

and

$$a_1 = \frac{x_k f_{k-1} - x_{k-1} f_k}{x_k - x_{k-1}}$$

From 1.5.1, we have $x = -\frac{a_1}{a_0}$. Hence substituting the values of a_0 and a_1 and writing x as x_{k+1} we get

$$x_{k+1} = -\frac{\left(\frac{x_k f_{k-1} - x_{k-1} f_k}{x_k - x_{k-1}}\right)}{\left(\frac{f_k - f_{k-1}}{x_k - x_{k-1}}\right)}$$

That is

$$x_{k+1} = \frac{f_k x_{k-1} - f_{k-1} x_k}{f_k - f_{k-1}} \qquad (1.5.4)$$

which can be expressed as

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f_k - f_{k-1}} f_k, \qquad k = 1, 2, \cdots \qquad (1.5.5)$$

This method is known as Secant Method or Chord Method.

Example 1.5.1. Using Secant Method, find the root of $f(x) = cosx - xe^x = 0$ taking the initial approximations as 0 and 1.

Solution: Here $x_0 = 0$, $f(x_0) = 1$ and $x_1 = 1$, $f(x_1) = -2.17798$. Using Secant formula

$$x_{k+1} = x_k - \frac{x_k - x_{k+1}}{f_k - f_{k+1}} f_k$$

we get

$$x_2 = x_1 - \frac{x_1 - x_0}{f_1 - f_0} f_1 = 0.314665$$

and $f(x_2) = 0.519872$. Then

$$x_3 = x_2 - \frac{x_2 - x_1}{f_2 - f_1} f_2 = 0.44674$$

and $f(x_3) = 0.2036$. Similarly, we get $x_4 = 0.531706$ and $x_5 = 0.51691$. Hence the approximate root is 0.51691.

1.5.2 Rate of Convergence

Let $f(x)$ be a continuous function. Then the Secant method to find the approximate root of $f(x) = 0$ is given by equation 1.5.5. Let ζ be the exact root of $f(x)$. Then we write x_{k+1}, x_k, x_{k-1} in terms of ζ as

$$x_{k+1} = \zeta + \epsilon_{k+1}, x_k = \zeta + \epsilon_k, x_{k-1} = \zeta + \epsilon_{k-1}$$

where ϵ_i is the error in the i^{th} iteration.

Hence, $f_k = f(x_k) = f(\zeta + \epsilon_k), f_{k-1} = f(x_{k-1}) = f(\zeta + \epsilon_{k-1})$. Thus, equation 1.5.5 becomes

$$\zeta + \epsilon_{k+1} = \zeta + \epsilon_k - \frac{\zeta + \epsilon_k - (\zeta + \epsilon_{k-1})}{f(\zeta + \epsilon_k) - f(\zeta + \epsilon_{k-1})} f(\zeta + \epsilon_k)$$

$$\epsilon_{k+1} = \epsilon_k - \frac{\zeta + \epsilon_k - (\zeta + \epsilon_{k-1})}{f(\zeta + \epsilon_k) - f(\zeta + \epsilon_{k-1})} f(\zeta + \epsilon_k)$$

Applying Taylor expansion on $f(\zeta + \epsilon_k)$ and $f(\zeta + \epsilon_{k-1})$ we get

$$\epsilon_{k+1} = \epsilon_k - \frac{(\epsilon_k - \epsilon_{k-1}) \left[f(\zeta) + \epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots \right]}{\left[f(\zeta) + \epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots \right] - \left[f(\zeta) + \epsilon_{k-1} f'(\zeta) + \frac{1}{2!}\epsilon_{k-1}^2 f''(\zeta) + \cdots \right]}$$

Since ζ is an exact root of $f(x) = 0$ we get

$$\begin{aligned}
\epsilon_{k+1} &= \epsilon_k - \frac{(\epsilon_k - \epsilon_{k-1})\left[\epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots\right]}{\left[\epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots\right] - \left[\epsilon_{k-1} f'(\zeta) + \frac{1}{2!}\epsilon_{k-1}^2 f''(\zeta) + \cdots\right]} \\
&= \epsilon_k - \frac{(\epsilon_k - \epsilon_{k-1})\left[\epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots\right]}{(\epsilon_k - \epsilon_{k-1}) f'(\zeta) + \frac{1}{2!}(\epsilon_k^2 - \epsilon_{k-1}^2) f''(\zeta) + \cdots} \\
&= \epsilon_k - \left[\frac{\epsilon_k f'(\zeta) + \frac{1}{2!}\epsilon_k^2 f''(\zeta) + \cdots}{f'(\zeta) + \frac{1}{2!}(\epsilon_k + \epsilon_{k-1}) f''(\zeta) + \cdots}\right] \\
&= \epsilon_k - \left[\frac{\epsilon_k + \frac{1}{2!}\frac{\epsilon_k^2 f''(\zeta)}{f'(\zeta)} + \cdots}{1 + \frac{(\epsilon_k + \epsilon_{k-1}) f''(\zeta)}{2! f'(\zeta)} + \cdots}\right] \\
&= \epsilon_k - \left[\epsilon_k + \frac{1}{2!}\frac{\epsilon_k^2 f''(\zeta)}{f'(\zeta)} + \cdots\right]\left[1 + \frac{(\epsilon_k + \epsilon_{k-1}) f''(\zeta)}{2! f'(\zeta)} + \cdots\right]^{-1} \\
&= \epsilon_k - \left[\epsilon_k + \frac{1}{2!}\frac{\epsilon_k^2 f''(\zeta)}{f'(\zeta)} + \cdots\right]\left[1 - \frac{(\epsilon_k + \epsilon_{k-1}) f''(\zeta)}{2! f'(\zeta)} + \cdots\right]
\end{aligned}$$

Neglecting the higher powers of $_k$ andϵ_{k-1}, we get

$$\epsilon_{k+1} = \frac{\epsilon_k \epsilon_{k-1} f''(\zeta)}{2! f'(\zeta)}$$

$$\text{Hence } \epsilon_{k+1} = C\epsilon_k \epsilon_{k-1} \qquad (1.5.6)$$

where $C = \frac{f''(\zeta)}{2! f'(\zeta)}$.

Considering the general equation of rate of convergence, we have

$$\epsilon_{k+1} = A\epsilon_k^p \qquad (1.5.7)$$

which implies

$$\epsilon_k = A\epsilon_{k-1}^p.$$

Then

$$\epsilon_{k-1} = A^{\frac{-1}{p}} \epsilon_k^{\frac{1}{p}}$$

Substituting the value of ϵ_{k-1} in 1.5.6 we get

$$\epsilon_{k+1} = C\epsilon_k A^{\frac{-1}{p}} \epsilon_k^{\frac{1}{p}}$$

From equation 1.5.7

$$\epsilon_{k+1} = A\epsilon_k^p = C\epsilon_k^{1+\frac{1}{p}} A^{\frac{-1}{p}} \qquad (1.5.8)$$

Comparing the powers of ϵ_k we get

$$p = 1 + \frac{1}{p}$$

which implies

$$p = \frac{1 \pm \sqrt{5}}{2}$$

Neglecting the negative value of p, we get the rate of convergence of Secant method as $\frac{1+\sqrt{5}}{2} \approx 1.618$. On comparing the constants of 1.5.8 we get,

$$A = C^{\frac{p}{1+p}}.$$

Thus, $\epsilon_{k+1} = C^{\frac{p}{1+p}} \epsilon_k^p$ where $p = \frac{1+\sqrt{5}}{2}$ and $C = \frac{f''(\zeta)}{2!f'(\zeta)}$.

1.6 Geometrical Interpretation of Secant and Regula Falsi Method

Geometrically, in Secant and Regula Falsi method we replace the function $f(x)$ by a chord passing through (x_k, f_k) and (x_{k-1}, f_{k-1}). We take the next root approximation as the point of intersection of the chord with the x- axis.

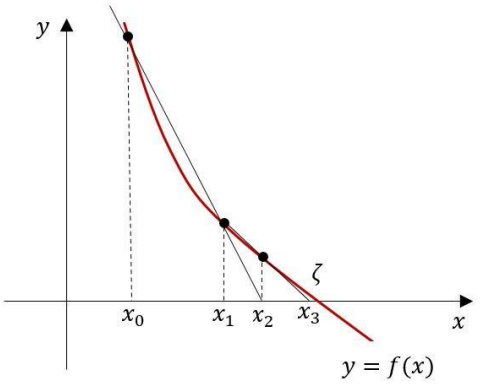


Figure 3: Secant Method

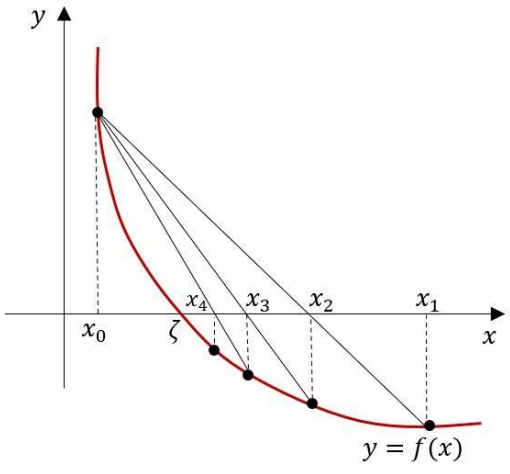


Figure 4: Regula Falsi Method

1.7 Summary

In this chapter, iteration methods to find the approximate roots of an equation is discussed.

The concepts of simple roots, multiple roots, algebraic and transcendental equations are discussed.

Intermediate value property to find the initial approximations to the root is discussed. Bisection method, Newton-Raphson method, Regula Falsi method and Secant method to find the approximate root of an equation is discussed.

Geometrical interpretation and rate of convergence of each method are discussed.

1.8 Exercise

1. Perform four iterations of bisection method to obtain a root of $f(x) = cosx - xe^x$.

2. Determine the initial approximation to find the smallest positive root for $f(x) = x^4 - 3x^2 + x - 10$ and find the root correct to five decimal places by Newton Raphson Method.

3. Perform four iterations of Newton Raphson method to obtain the approximate value of $17^{\frac{1}{3}}$ starting with the initial approximation $x_0 = 2$.

4. Using Regula Falsi Method, find the root of $f(x) = cosx - xe^x = 0$ taking the initial approximations as 0 and 1.

5. Perform three iterations of Secant Method for $f(x) = x^3 - 5x + 1$ such that the root lies in the interval $(0,1)$.

6. For $f(x) = x^4 - x - 10$ determine the initial approximations to find the smallest positive root correct up to three decimal places using Newton Raphson method, Secant Method and Regula Falsi Method.

7. Show that the Newton-Raphson method leads to the recurrence

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right)$$

to find the square-root of a.

8. For $f(x) = x - e^{-x} = 0$ determine the initial approximation to find the smallest positive root. Find the root correct to three decimal places using Regula Falsi and Secant method.

9. Find the approximate root correct upto three decimal places for $f(x) = \cos x - xe^{x}$ using Newton- Raphson Method with initial approximation $x_0 = 1$.

❖❖❖

UNIT 1

2

INTERPOLATION

Unit Structure

2.0 Objectives

This chapter will enable the learner to:

- Understand the concepts of interpolation and interpolating polynomial. Prove the existence and uniqueness of an interpolating polynomial.
- To find interpolating polynomial using Lagrange method and Newton Divided Difference method.

- To understand the concepts of finite difference operators and to relate between difference operators and divided differences.
- To find interpolating polynomial using finite differences using Newton Forward Difference and Backward Difference Interpolation.

2.1 Introduction

If we have a set of values of a function $y = f(x)$ as follows:

x	x_0	x_1	x_2	$\cdots$	x_n
y	y_0	y_1	y_2	$\cdots$	y_n

then the process of finding the value of $f(x)$ corresponding to any value of $x = x_i$ between x_0 and x_n is called interpolation. If the function $f(x)$ is explicitly known, then the value of $f(x)$ corresponding to any value of $x = x_i$ can be found easily. But, if the explicit form of the function is not known then finding the value of $f(x)$ is not easy. In this case, we approximate the function by simpler functions like polynomials which assumes the same values as those of $f(x)$ at the given points x_0, x_1, $x_2, \cdots, x_n$.

Definition 2.1.1 (Interpolating Polynomial). A polynomial $P(x)$ is said to be an interpolating polynomial of a function $f(x)$ if the values of $P(x)$ and/or its certain order derivatives coincides with those of $f(x)$ for given values of x.

If we know the values of $f(x)$ at $n + 1$ distinct points say $x_0 < x_1 < x_2 < \cdots < x_n$, then interpolation is the process of finding a polynomial $P(x)$ such that

(a) $P(x_i) = f(x_i), \quad i = 0,1,2,\cdots,n$

or

(a) $P(x_i) = f(x_i), \quad i = 0,1,2,\cdots,n$

(b) $P'(x_i) = f'(x_i), \quad i = 0,1,2,\cdots,n$

2.1.1 Existence and Uniqueness of Interpolating Polynomial

Theorem 2.1.1.1. Let $f(x)$ be a continuous function on $[a,b]$ and let $a = x_0 < x_1 < x_2 < \cdots < x_n = b$ be the $n + 1$ distinct points such that the value of $f(x)$ is known at these points. Then there exist a unique polynomial $P(x)$ such that $P(x_i) = f(x_i)$, for $i = 0,1,2,\cdots,n$.

Proof. We intend to find a polynomial $P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ such that $P(x_i) = f(x_i)$, for $i = 0,1,2,\cdots,n$. That is

$$\begin{aligned} a_0 + a_1x_0 + a_2x_0^2 + \cdots + a_nx_0^n &= f(x_0) \\ a_0 + a_1x_1 + a_2x_1^2 + \cdots + a_nx_1^n &= f(x_1) \\ a_0 + a_1x_2 + a_2x_2^2 + \cdots + a_nx_2^n &= f(x_2) \\ \vdots \qquad\qquad & \qquad \vdots \\ a_0 + a_1x_n + a_2x_n^2 + \cdots + a_nx_n^n &= f(x_n) \end{aligned} \tag{2.1.1.1}$$

Then the polynomial *P(x)* exists only if the system of equations 2.1.1.1 has a unique solution. That is, the polynomial *P(x)* exists only if the Vandermonde's determinant is non-zero (in other words, *P(x)* exists only if the determinant of the co-efficient matrix is non-zero). Let

$$V(x_0, x_1, \cdots, x_{n-1}, x_n) = \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix}$$

By the properties of determinants, we get

$$V(x_0, x_1, \cdots, x_{n-1}, x_n) = \prod_{\substack{i,j=0 \\ i>j}}^{n} (x_i - x_j)$$

Since each x_i's are distinct, we get $\prod_{\substack{i,j=0 \\ i>j}}^{n} (x_i - x_j) \neq 0$ and hence the system of equations 5.1.1.1 has a unique solution.

To prove the uniqueness of $P(x)$, let there exists another polynomial $Q(x)$ such that $Q(x_i) = f(x_i)$, for $i = 0, 1, 2, \cdots, n$.

Let

$$R(x) = P(x) - Q(x).$$

As $P(x)$ and $Q(x)$ are polynomials of degree $\leq n$, we get that $R(x)$ is also a polynomial of degree $\leq n$. Also, $R(x_i) = 0$, for $i = 0, 1, 2, \cdots, n$.

That is $R(x)$ is a polynomial of degree $\leq n$ with $n + 1$ distinct roots $x_0, x_1, x_2, \cdots, x_n$ which implies that $R(x) = 0$. Hence $P(x) = Q(x)$. □

2.2 Lagrange Interpolation

2.2.1 Linear Interpolation

For linear interpolation, $n = 1$ and we find a 1-degree polynomial

$$P_1(x) = a_1x + a_0$$

such that

$$f(x_0) = P_1(x_0) = a_1x_0 + a_0$$

and

$$f(x_1) = P_1(x_1) = a_1x_1 + a_0.$$

We eliminate a_0 and a_1 to obtain $P_1(x)$ as follows:

$$\begin{vmatrix} P_1(x) & x & 1 \\ f(x_0) & x_0 & 1 \\ f(x_1) & x_1 & 1 \end{vmatrix} = 0$$

On simplifying, we get

$P_1(x)(x_0 - x_1) - f(x_0)(x - x_1) + f(x_1)(x - x_0) = 0$ and hence

$$\begin{aligned} P_1(x) &= \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1) \\ &= l_0(x) f(x_0) + l_1(x) f(x_1) \end{aligned} \qquad (2.2.1.2)$$

where $l_0(x) = \frac{(x-x_1)}{(x_0-x_1)}$ and $l_1(x) = \frac{(x-x_0)}{(x_1-x_0)}$ are called the Lagrange Fundamental Polynomials which satisfies the condition $l_0(x) + l_1(x) = 1$.

Example 2.2.1. Find the unique polynomial $P(x)$ such that $P(2) = 1.5$, $P(5) = 4$ using Lagrange interpolation.

Solution: By Lagrange interpolation formula,

$$\begin{aligned} P_1(x) &= \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1) \\ &= \frac{(x - 5)}{(2 - 5)} 1.5 + \frac{(x - 2)}{(5 - 2)} 4 \\ &= \frac{5x - 1}{6} \end{aligned}$$

2.2.2 Quadratic Interpolation

For quadratic interpolation, $n = 2$ and we find a 2-degree polynomial

$$P_2(x) = a_2x^2 + a_1x + a_0$$

such that

$$f(x_0) = P_2(x_0) = a_2x_0^2 + a_1x_0 + a_0$$

$$f(x_1) = P2(x_1) = a_2x_1^2 + a_1x_1 + a_0.$$

$$f(x_2) = P_2(x_2) = a_2x_2^2 + a_1x_2 + a_0$$

We eliminate a_0, a_1 and a_2 to obtain $P_2(x)$ as follows:

$$\begin{vmatrix} P_2(x) & 1 & x & x^2 \\ f(x_0) & 1 & x_0 & x_0^2 \\ f(x_1) & 1 & x_1 & x_1^2 \\ f(x_2) & 1 & x_2 & x_2^2 \end{vmatrix} = 0$$

On simplifying, we get

$$\begin{aligned} P_2(x) =& \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f(x_1) \\ &+ \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}f(x_2) \\ =& l_0(x)f(x_0) + l_1(x)f(x_1) + l_2(x)f(x_2) \end{aligned}$$

where $l_0(x) + l_1(x) + l_2(x) = 1$.

Example 2.2.2. Find the unique polynomial $P(x)$ such that $P(3) = 1$, $P(4) = 2$ and $P(5) = 4$ using Lagrange interpolation.

Solution: By Lagrange interpolation formula,

$$\begin{aligned} P_2(x) =& \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f(x_1) \\ &+ \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}f(x_2) \\ =& \frac{(x-4)(x-5)}{(3-4)(3-5)}(1) + \frac{(x-3)(x-5)}{(4-3)(4-5)}(2) + \frac{(x-3)(x-4)}{(5-3)(5-4)}(4) \\ =& \frac{x^2-5x+8}{2} \end{aligned}$$

2.2.3 Higher Order Interpolation

The Lagrange Interpolating polynomial $P(x)$ of degree n for given $n + 1$ distinct points $a = x_0 < x_1 < x_2 < \cdots < x_n = b$ is given by

$$P_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i) \quad (2.2.3.1)$$

where,

$$l_i(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}$$

for $i = 0, 1, \cdots, n$.

2.3 Newton Divided Difference Interpolation

2.3.1 Linear Interpolation

For linear interpolation, $n = 1$ and we find a 1-degree polynomial

$$P_1(x) = a_1 x + a_0$$

such that

$$f(x_0) = P_1(x_0) = a_1 x_0 + a_0$$

and

$$f(x_1) = P_1(x_1) = a_1 x_1 + a_0.$$

We eliminate a_0 and a_1 to obtain $P_1(x)$ as follows:

$$\begin{vmatrix} P_1(x) & x & 1 \\ f(x_0) & x_0 & 1 \\ f(x_1) & x_1 & 1 \end{vmatrix} = 0$$

On simplifying, we get

$$P_1(x) = f(x_0) + (x - x_0) \frac{f(x_1) - f(x_0)}{(x_1 - x_0)}$$

Let $f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{(x_1 - x_0)}$, then

$$P_1(x) = f(x_0) + (x - x_0) f[x_0, x_1] \qquad (2.3.1.1)$$

The ratio $f[x_0, x_1]$ is called the first divided difference of $f(x)$ relative to x_0 and x_1. The equation 2.3.1.1 is called the linear Newton interpolating polynomial with divided differences.

Example 2.3.1. Find the unique polynomial $P(x)$ such that $P(2) = 1.5$, $P(5) = 4$ using Newton divided difference interpolation.

Solution: Here

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{5}{6}$$

Hence, by Newton divided difference interpolation,

$$\begin{aligned} P_1(x) =& f(x_0) + (x - x_0) f[x_0, x_1] \\ =& 1.5 + \frac{5}{6}(x - 2) \\ =& \frac{5x - 1}{6} \end{aligned}$$

2.3.2 Higher Order Interpolation

We define higher order divided differences as

$$\begin{aligned} f[x_0, x_1, x_2] =& \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} \\ =& \frac{1}{x_2 - x_0} \left[\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0} \right] \\ =& \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)} \end{aligned}$$

Hence, in general we have

$$f[x_0, x_1, x_2, \cdots, x_{k-1}, x_k] = \frac{f[x_1, x_2, \cdots, x_k] - f[x_0, x_1, x_2, \cdots, x_{k-1}]}{x_k - x_0}$$

for $k = 3,4,\cdots,n$ and in terms of function values, we have

$$f[x_0, x_1, x_2, \cdots, x_{k-1}, x_k] = \sum_{i=0}^{n} \frac{f(x_i)}{\prod\limits_{\substack{j=0 \\ i \neq j}}^{n} (x_i - x_j)}$$

Then the Newton's Divided Difference interpolating polynomial is given by

$$\begin{aligned} P_n(x) =& f[x_0] + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \cdots \\ &+ (x - x_0)(x - x_1) \cdots (x - x_{n-1}) f[x_0, x_1, \cdots, x_n] \end{aligned} \qquad (2.3.2.1)$$

Example 2.3.2. Construct the divided difference table for the given data and hence find the interpolating polynomial.

x	0.5	1.5	3.0	5.0	6.5	8.0
$f(x)$	1.625	5.875	31.000	131.000	282.125	521.000

Solution: The divided difference table will be as follows:

x	$f(x)$	I order d.d.	II order d.d.	III order d.d.	IV order d.d.
0.5	1.625				
		4.25			
1.5	5.875		5		
		16.75		1	
3.0	31.000		9.5		0
		50.00		1	
5.0	131.000		14.5		0
		100.75		1	
6.5	282.125		19.5		
		159.25			
8.0	521.000				

From the table, as the fourth divided differences are zero, the interpolating polynomial will be given as

$$\begin{aligned} P_3(x) &= f[x_0] + (x - x_0)f[x_0,x_1] + (x - x_0)(x - x_1)f[x_0,x_1,x_2] \\ &\quad + (x - x_0)(x - x_1)(x - x_2)f[x_0,x_1,x_2,x_3] \\ &= 1.625 + (x - 0.5)(4.25) + 5(x - 0.5)(x - 1.5) \\ &\quad + 1(x - 0.5)(x - 1.5)(x - 3) \\ &= x^3 + x + 1 \end{aligned}$$

2.4 Finite Difference Operators

Consider n+1 equally spaced points $x_0,x_1,\cdots,x_n$ such that $x_i = x_0+ih, i = 0,1,\cdots,n$. We define the following operators:

1. Shift Operator (E): $E(f(x_i)) = f(x_i + h)$.

2. Forward Difference Operator (Δ): $\Delta(f(x_i)) = f(x_i + h) - f(x_i)$.

3. Backward Difference Operator (∇): $\nabla(f(x_i)) = f(x_i) - f(x_i - h)$.

4. Central Difference Operator (δ $\delta(f(x_i)) = f(x_i + \frac{h}{2}) - f(x_i - \frac{h}{2})$.

5. Average Operator (μ $\mu(f(x_i)) = \frac{1}{2}\left[f(x_i + \frac{h}{2}) + f(x_i - \frac{h}{2})\right]$.

Some properties of these operators:

$\Delta f_i = \nabla f_{i+1} = \delta f_{i+\frac{1}{2}}$.

$\Delta = E - I$

$\nabla = I - E^{-1}$

$\delta = E^{1/2} - E^{-1/2}$

$\mu = \frac{1}{2}(E^{1/2} + E^{-1/2})$

Now we write the Newton's divided differences in terms of forward and backward difference operators. Consider

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{(x_1 - x_0)}$$

As $x_i = x_0 + ih$, $i = 0, 1, \cdots, n$, we get $x_1 - x_0 = h$ and hence

$$f[x_0, x_1] = \frac{1}{h}\Delta f_0$$

Now we consider

$$\begin{aligned} f[x_0, x_1, x_2] &= \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} \\ &= \frac{\frac{1}{h}\Delta f_1 - \frac{1}{h}\Delta f_0}{2h} \\ &= \frac{1}{2!h^2}\Delta^2 f_0 \end{aligned}$$

Thus, by induction we have

$$f[x_0, x_1, \cdots, x_n] = \frac{1}{n!h^n}\Delta^n f_0.$$

Similarly, for backward difference operator, consider

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{(x_1 - x_0)} = \frac{1}{h}\nabla f_1$$

$$\begin{aligned} f[x_0, x_1, x_2] =& \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} \\ =& \frac{\frac{1}{h}\nabla f_2 - \frac{1}{h}\nabla f_1}{2h} \\ =& \frac{1}{2!h^2}\nabla^2 f_2 \end{aligned}$$

Thus, by induction we have

$$f[x_0, x_1, \cdots, x_n] = \frac{1}{n!h^n}\nabla^n f_n.$$

2.5 Interpolating polynomials using finite differences operators

2.5.1 Newton Forward Difference Interpolation

Newton's forward difference interpolating polynomial is obtained by substituting the divided differences in 2.3.2.1 with the forward differences. That is

$$\begin{aligned} P_n(x) =& f[x_0] + (x - x_0)f[x_0, x_1] + \cdots \\ & + (x - x_0)(x - x_1)\cdots(x - x_{n-1})f[x_0, x_1, \cdots, x_n] \\ =& f(x_0) + \frac{(x - x_0)}{h}\Delta f_0 + \frac{(x - x_0)(x - x_1)}{2!h^2}\Delta^2 f_0 + \cdots \\ & + \frac{(x - x_0)(x - x_1)\cdots(x - x_{n-1})}{n!h^n}\Delta^n f_0 \end{aligned} \qquad (2.5.1.1)$$

Let$u = \frac{(x-x_0)}{h}$, then 2.5.1.1 can be written as

$$\begin{aligned} P_n(x) =& P_n(x_0 + hu) \\ =& f(x_0) + u\Delta f(x_0) + \frac{u(u-1)}{2!}\Delta^2 f(x_0) + \cdots \\ & + \frac{u(u-1)\cdots(u-n+1)}{n!}\Delta^n f(x_0) \end{aligned} \qquad (2.5.1.2)$$

2.5.2 Newton Backward Difference Interpolation

Newton interpolation with divided differences can be expressed in terms of backward differences by evaluating the differences at the end point x_n. Hence, we, write

$$f(x) = f(x_n + \frac{x - x_n}{h}h) = f(x_n + hu) = E^u f(x_n) = (1 - \nabla)^{-u} f(x_n)$$

where$u = \frac{x-x_n}{h}$.

Expanding $(1-\nabla)^{-u}$ in binomial series, we get

$$f(x) = f(x_n) + u\nabla f(x_n) + \frac{u(u+1)}{2!}\nabla^2 f(x_n) + \cdots + \frac{u(u+1)\cdots(u+n-1)}{n!}\nabla^n f(x_n) + \cdots$$

Neglecting the higher order differences, we get the interpolating polynomial as

$$P_n(x) = P_n(x_n + hu)$$

$$= f(x_n) + u\nabla f(x_n) + \frac{u(u+1)}{2!}\nabla^2 f(x_n) + \cdots + \frac{u(u+1)\cdots(u+n-1)}{n!}\nabla^n f(x_n) \qquad (2.5.2.1)$$

Example 2.5. Obtain the Newton's forward and backward difference interpolating polynomial for the given data

x	0.1	0.2	0.3	0.4	0.5
$f(x)$	1.40	1.56	1.76	2.00	2.28

Solution: The difference table will be as follows:

x	$f(x)$	I order d.d.	II order d.d.	III order d.d.	IV order d.d.
0.1	1.40				
		0.16			
0.2	1.56		0.04		
		0.20		0	
0.3	1.76		0.04		0
		0.24		0	
0.4	2.00		0.04		
		0.28			
0.5	2.28				

From the table, as the third differences onwards are zero, the interpolating polynomial using Newton's forward differences will be given as

$$P_2(x) = 1.4 + \frac{0.16}{0.1}(x-0.1) + \frac{0.04}{0.01}\frac{(x-0.1)(x-0.2)}{2}$$
$$= 2x^2 + x + 1.28$$

and the Newton's backward difference interpolating polynomial will be given as

$$P_2(x) = 2.28 + \frac{0.28}{0.1}(x - 0.5) + \frac{0.04}{0.01}\frac{(x - 0.5)(x - 0.4)}{2}$$
$$= 2x^2 + x + 1.28$$

2.6 Summary

In this chapter, interpolation methods to approximate a function by a family of simpler functions like polynomials is discussed.

Lagrange Interpolation method and Newton Divided Difference Interpolation method are discussed.

Difference operators namely shift operator, forward, backward and central difference operators and average operator are discussed.

The relation between divided difference and the forward and backward difference operators are discussed and hence the Newton's divided difference interpolation is expressed in terms of forward and backward difference operators.

2.7 Exercises

1. Given $f(2) = 4$ and $f(2.5) = 5.5$, find the linear interpolating polynomial using Lagrange interpolation and Newton divided difference interpolation.

2. Using the data $\sin 0.1 = 0.09983$ and $\sin 0.2 = 0.19867$, find an approximate value of $\sin 0.15$ using Lagrange interpolation.

3. Using Newton divided difference interpolation, find the unique polynomial of degree 2 such that $f(0) = 1, f(1) = 3$ and $f(3) = 55$.

4. Calculate the n^{th} divided difference of $\frac{1}{x}$ for points $x_0, x_1, \cdots, x_n$.

5. Show that $\delta = \nabla(1 - \nabla)^{-1/2}$ and $\mu = \left[1 + \frac{\delta^2}{4}\right]^{1/2}$.

6. For the given data, find the Newton's forward and backward difference polynomials.

x	0	0.1	0.2	0.3	0.4	0.5
$f(x)$	−1.5	−1.27	−0.98	−0.63	−0.22	0.25

7. Calculate $f[x_1, x_2, x_3, x_4]$ $\quad f(x) = \frac{1}{x^2}$.

8. If $f(x) = e^{ax}$, show that $\Delta^n f(x) = (e^{ah} - 1)^n e^{ax}$.

9. Using the Newton's backward difference interpolation, construct the interpolating polynomial that fits data.

x	0.1	0.3	0.5	0.7	0.9	1.1
$f(x)$	−1.699	−1.073	−0.375	0.443	1.429	2.631

10. Find the interpolating polynomial that fits the data as follows:

x	−2	−1		0	1	3	4
$f(x)$	9	16		17	18	44	81

❖❖❖

UNIT 2

3

SOLUTION OF SIMULTANEOUS ALGEBRAIC EQUATIONS

Unit Structure

3.0 Objectives

Student will be able to understand the following from the Chapter:

Method to represent equations in Matrix Form.

Rules of Elementary Transformation.

Application of Dynamic Iteration Method.

3.1 Introduction

An equation is an expression having two equal sides, which are separated by an *equal* to sign. For Example: 9 + 4 = 13

Here, a mathematical operation has been performed in which 4 and 9 has been added and the result 13 has been written alongside with an **equal** to sign.

Similarly, an **Algebraic** equation is a mathematical expression having two equal sides separated by an *equal to* sign, in which one side the expression is formulated by a set of variables and on the other side there is a constant value. For example $2x^2 + 5y^3 = 9$,

Here, the set of variables x and y has been used to provide a unique pair of values whose sum will be equal to 9.

If the **powers** (degrees) of the variables used is 1, then these algebraic equations are known as Linear Algebraic equation. For example: $2x + 3y + 5z = 8$.

It may happen, that there are more than equations,where there will be at-lest one unique pair which will satisfy all the Algebraic equations. The procedure to find these unique pairs are known as Solutions of Simultaneous Algebraic Equations.

There are two types of methods:

1. Gauss-Jordan Method
2. Gauss-Seidel Method

3.2 Gauss-Jordan Method

Gauss Jordan Method is an algorithm used to find the solution of simultaneous equations. The algorithm uses **Matrix Approach** to determine the solution.

The method requires elementary transformation or elimination using ow operations. Hence, it is also known as **Gauss-Elimination Method**.

Steps of Algorithm:

i Represent the set of equation in the following format:

$$A \times X = B$$

where,

A: Coefficient Matrix

X: Variable Matrix B: Constant Matrix

Example:

Convert the following equations in Matrix format:

$3x + 5y = 12$

$2x + y = 1$

The Matrix representation is:

In the above set of equations, the coefficients are the values which are written along-with the variables.

The Constant matrix are the values which are written after the equal to sign.
Hence, the coefficient matrix is given as:

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

Hence, the variable matrix is given as:

$$\begin{bmatrix} x \\ y \end{bmatrix}$$

And, the Constant matrix is:

$$\begin{bmatrix} 12 \\ 1 \end{bmatrix}$$

ii Temporarily Combine the Coefficient *A* and Constant *B* Matrices in the following format.

$$[A : B]$$

iii Perform row Transformation considering following Do's and Don't:

Allowed	**Not Allowed**
Swapping of Rows $Ra \leftrightarrow Rb$ where a 6= b	Swapping between Row and Column $Ra \leftrightarrow Cb$
Mathematical Operations between Rows allowed: *Addition, Subtraction.* $Ra \leftrightarrow Ra \pm Rb$ *Mathematical Operations with constant value allowed:* *Multiplication, Division.* $R_a \leftrightarrow R_a \times k$ $R_a \leftrightarrow \frac{R_a}{k}$	*Mathematical Operations between Rows not allowed:* *Multiplication, Division.* $R_a \leftrightarrow R_a \times R_b$ $R_a \leftrightarrow \frac{R_a}{R_b}$ *Mathematical Operations with constant value allowed:* *Addition, Subtraction.* $Ra \leftrightarrow Ra \pm k$

Row Transformation is done to convert the **Coefficient Matrix** *A* to **Unit Matrix** of same dimension as that of *A*.

3.2.1 Solved Examples:

i. Solve the system-

$6x + y + z = 20$ $x + 4y - z = 6$ $x - y + 5z = 7$

using Gauss-Jordan Method

Sol. Given:

$6x + y + z = 20$ $x + 4y - z = 6$ $x - y + 5z = 7$

The matrix representation is-

$$\begin{bmatrix} 6 & 1 & 1 \\ 1 & 4 & -1 \\ 1 & -1 & 5 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 20 \\ 6 \\ 7 \end{bmatrix}$$

Followed by the echolon form, formed by combining Coefficient Matrix and Constant Matrix.

$$\begin{bmatrix} 6 & 1 & 1 & : & 20 \\ 1 & 4 & -1 & : & 6 \\ 1 & -1 & 5 & : & 7 \end{bmatrix}$$

Perform elementary ow transformation in the above matrix to convert

matrix *A* to the following:

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

$$\begin{bmatrix} 6 & 1 & 1 & : & 20 \\ 1 & 4 & -1 & : & 6 \\ 1 & -1 & 5 & : & 7 \end{bmatrix}$$

$R_1 \leftrightarrow R_2$

$$\begin{bmatrix} 1 & 4 & -1 & : & 6 \\ 6 & 1 & 1 & : & 20 \\ 1 & -1 & 5 & : & 7 \end{bmatrix}$$

$R_2 \leftrightarrow R_2 - 6R_1$

$$\begin{bmatrix} 1 & 4 & -1 & : & 6 \\ 0 & -23 & 7 & : & -16 \\ 1 & -1 & 5 & : & 7 \end{bmatrix}$$

$R_3 \leftrightarrow R_3 - R_1$

$$\begin{bmatrix} 1 & 4 & -1 & : & 6 \\ 0 & -23 & 7 & : & -16 \\ 0 & -5 & 6 & : & 1 \end{bmatrix}$$

$$R_3 \leftrightarrow R_3 - R_2$$

$$\begin{bmatrix} 1 & 4 & -1 & : & 6 \\ 0 & -23 & 7 & : & -16 \\ 0 & 18 & -1 & : & 17 \end{bmatrix}$$

$$R_2 \leftrightarrow \frac{R_2}{-23}$$

$$\begin{bmatrix} 1 & 4 & -1 & : & 6 \\ 0 & 1 & \frac{-7}{23} & : & \frac{16}{23} \\ 0 & -18 & 1 & : & -17 \end{bmatrix}$$

$$R_3 \leftrightarrow R_3 - 18R_2$$

$$\begin{bmatrix} 1 & 4 & -1 & : & 6 \\ 0 & 1 & \frac{-7}{23} & : & \frac{16}{23} \\ 0 & 0 & \frac{103}{23} & : & \frac{103}{23} \end{bmatrix}$$

$$R_2 \leftrightarrow \frac{23}{103} \times R_3$$

$$\begin{bmatrix} 1 & 4 & -1 & : & 6 \\ 0 & 1 & \frac{-7}{23} & : & \frac{16}{23} \\ 0 & 0 & 1 & : & 1 \end{bmatrix}$$

$$R_1 \leftrightarrow R_1 - 4R_2$$

$$\begin{bmatrix} 1 & 0 & \frac{5}{23} & : & \frac{74}{23} \\ 0 & 1 & \frac{-7}{23} & : & \frac{16}{23} \\ 0 & 0 & 1 & : & 1 \end{bmatrix}$$

$$R_2 \leftrightarrow R_2 + \frac{7}{23} \times R_2$$

$$\begin{bmatrix} 1 & 0 & \frac{5}{23} & : & \frac{74}{23} \\ 0 & 1 & 0 & : & 1 \\ 0 & 0 & 1 & : & 1 \end{bmatrix}$$

$$R_1 \leftrightarrow R_1 - \frac{5}{23} \times R_3$$

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

The solution of the equations are: $x = 3$; $y = 1$ and $z = 1$.

ii. Solve the system-

$2x + y + z = 10$ $3x + 2y + 3z = 18$ $x + 4y + 9z = 16$

using Gauss-Jordan Method

Sol. Given:

$2x + y + z = 10$ $3x + 2y + 3z = 18$ $x + 4y + 9z = 16$

The matrix representation is-

$$\begin{bmatrix} 2 & 1 & 1 \\ 3 & 2 & 3 \\ 1 & 4 & 9 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 10 \\ 18 \\ 16 \end{bmatrix}$$

Followed by the echolon form, formed by combining Coefficient Matrix and Constant Matrix.

$$\left[\begin{array}{ccc:c} 2 & 1 & 1 & 10 \\ 3 & 2 & 3 & 18 \\ 1 & 4 & 9 & 16 \end{array}\right]$$

Perform elementary ow transformation in the above matrix to convert matrix A to the following:

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

$$\left[\begin{array}{ccc:c} 2 & 1 & 1 & 10 \\ 3 & 2 & 3 & 18 \\ 1 & 4 & 9 & 16 \end{array}\right]$$

$$R_1 \leftrightarrow R_3$$

$$\left[\begin{array}{ccc:c} 1 & 4 & 9 & 16 \\ 3 & 2 & 3 & 18 \\ 2 & 1 & 1 & 10 \end{array}\right]$$

$$R_2 \leftrightarrow R_2 - 3R_1$$

$$\left[\begin{array}{ccc:c} 1 & 4 & 9 & 16 \\ 0 & -10 & -24 & -30 \\ 2 & 1 & 1 & 10 \end{array}\right]$$

$$R_3 \leftrightarrow R_3 - 2R_1$$

$$\begin{bmatrix} 1 & 4 & 9 & : & 16 \\ 0 & -10 & -24 & : & -30 \\ 0 & -7 & -17 & : & -22 \end{bmatrix}$$

$$R_3 \leftrightarrow -R_3$$

$$\begin{bmatrix} 1 & 4 & 9 & : & 16 \\ 0 & -10 & -24 & : & -30 \\ 0 & 7 & 17 & : & 22 \end{bmatrix}$$

$$R_2 \leftrightarrow -R_2$$

$$\begin{bmatrix} 1 & 4 & 9 & : & 16 \\ 0 & 10 & 24 & : & 30 \\ 0 & 7 & 17 & : & 22 \end{bmatrix}$$

$$R_2 \leftrightarrow \frac{R_2}{10}$$

$$\begin{bmatrix} 1 & 4 & 9 & : & 16 \\ 0 & 1 & \frac{24}{10} & : & 3 \\ 0 & 7 & 17 & : & 22 \end{bmatrix}$$

$$R_3 \leftrightarrow R_3 - 7R_2$$

$$\begin{bmatrix} 1 & 4 & 9 & : & 16 \\ 0 & 1 & \frac{24}{10} & : & 3 \\ 0 & 0 & \frac{1}{5} & : & 1 \end{bmatrix}$$

$$R_1 \leftrightarrow R_1 - 4R_2$$

$$\begin{bmatrix} 1 & 0 & \frac{-3}{5} & : & 4 \\ 0 & 1 & \frac{24}{10} & : & 3 \\ 0 & 0 & \frac{1}{5} & : & 1 \end{bmatrix}$$

$$R_3 \leftrightarrow 5R_3$$

$$\begin{bmatrix} 1 & 0 & \frac{-3}{5} & : & 4 \\ 0 & 1 & \frac{24}{10} & : & 3 \\ 0 & 0 & 1 & : & 5 \end{bmatrix}$$

$$R_2 \leftrightarrow R_2 - \frac{24}{10}R_3$$

$$\begin{bmatrix} 1 & 0 & \frac{-3}{5} & : & 4 \\ 0 & 1 & 0 & : & -9 \\ 0 & 0 & 1 & : & 5 \end{bmatrix}$$

$$R_1 \leftrightarrow R_1 + \frac{3}{5}R_3$$

$$\begin{bmatrix} 1 & 0 & 0 & : & 7 \\ 0 & 1 & 0 & : & -9 \\ 0 & 0 & 1 & : & 5 \end{bmatrix}$$

The solution of the equations are: $x = 7$; $y = -9$ and $z = 5$.

3.3 Gauss-Seidel Method

Gauss Seidel Method uses Iterative methods to find the unique solution of Linear Algebraic equations. In this method, the present value of a variable depends on the past and present value of the other variables. This type of Iteration is known as *Dynamic Iteration Method*.

To achieve convergence of the values it is important to have **Diagonal Dominance**. In Diagonal Dominance, the first equation should have the highest coefficient among the set of x coefficients as well as it should be the highest coefficient within the same equation. Similarly, the second equation should have highest coefficient of y as well as among its own equation.

After ensuring Diagonal Dominance, the variable of each equation is represented as the function of other variables.

3.3.1 Solved Examples:

i. Solve the equation: $x + 4y - z = 6$ $6x + y + z = 20$ $x - y + 5z = 7$

by using Gauss-Seidel Method

Sol. Given:

$x + 4y - z = 6$ $6x + y + z = 20$ $x - y + 5z = 7$

On comparing the coefficient of x among the given set of equations. The maximum value is 6 which is present in the second equation. Considering the second equation, the maximum coefficient present among the variable is also 6 (Coefficient of x)

Hence, the first equation is: **6x+y+z=20 —(i)**

Similarly, among First and Third equation, the maximum value of the Coefficient of *y* is 4.

Hence, the second equation is: **x+4y-z=6—(ii)**

And, the third equation is: **x-y+5z=7—(iii)**

Now represent each variable as a function of other two variables like: Using equation 1:

$$x = \frac{20 - y - z}{6}$$

Using equation 2:

$$y = \frac{6 - x + z}{4}$$

Using equation 3:

$$z = \frac{7 + y - x}{5}$$

Consider the initial values of *x*, *y* and *z* as 0.

Now,

To implement the iteration the equations are re-written as:

$$x_n = \frac{20 - y_{n-1} - z_{n-1}}{6}$$
$$y_n = \frac{6 - x_n + z_{n-1}}{4}$$
$$z_n = \frac{7 + y_n - x_n}{5}$$

Where, x_n, y_n and z_n are the **Present** values of *x*, *y* and *z* respectively. x_{n-1}, y_{n-1} and z_{n-1} are the **Past** values of *x*, *y* and *z* respectively.

Means,

To calculate the Present value of *x*, we require Past values of *y* and *z*.

To calculate the Present value of *y*, we require Present value of *x* and Past value of *z*.

To calculate the Present value of *z*, we require Present values of *x* and *y*.

Hence the values of x, y and z will be:

i	x_n	y_n	z_n
0	0	0	0
1	3.33	0.6675	0.8675
2	3.0775	0.9475	0.974
3	3.0131	0.97412	0.992
4	3.0056	0.99665	0.9982
5	3.0035	0.99815	0.99893
6	3.0005	0.9996	0.99982
7	3.00009	0.99993	0.999668

On Approximation:

$x = 3$; $y = 1$ and $z = 1$ ii. Solve the equation:

$x_1 + 10x_2 + 4x_3 = 6$ $2x_1 + 10x_2 + 4x_3 = -15$

$9x_1 + 2x_2 + 4x_3 = 20$

by using Gauss-Seidel Method

Sol. Given:

$x_1 + 10x_2 + 4x_3 = 6$

$2x_1 - 4x_2 + 10x_3 = -15$

$9x_1 + 2x_2 + 4x_3 = 20$

On re-arranging to achieve the Diagonal Dominance:

$9x_1 + 2x_2 + 4x_3 = 20$ — (1) $x_1 + 10x_2 + 4x_3 = 6$ —(2)

$2x_1 - 4x_2 + 10x_3 = -15$ —(3)

Therefore,

$$x_{1_n} = \frac{20 - 2x_{2_{n-1}} - 4x_{3_{n-1}}}{9}$$ (From equation (1))

$$x_2 = \frac{6 - x_{1_n} - 4x_{3_{n-1}}}{10}$$ (From equation (2))

$$x_3 = \frac{-15 - 2x_{1_n} + 4x_{2_n}}{10}$$ (From equation (3)) Hence the values of x, y and z will be:

i	x_n	y_n	z_n
0	0	0	0
1	2.2222	0.3778	-1.7933
2	2.9353	1.0238	-1.6775
3	2.7403	0.99697	-1.6493
4	2.7337	0.98635	-1.6522
5	2.7373	0.98715	-1.6526
6	2.7373	0.98731	-1.6525
7	2.7373	0.9873	-1.6525

On Approximation:

$x = 2.7373$; $y = 0.9873$ and $z = -1.6525$

3.4 Summary

Linear Algebraic equations can be solved by using two methods.

- Gauss Seidel Method
- Gauss Jordan Method

Gauss Seidel Method uses iterative approach, following Diagonal Dominance principle.

Gauss Jordan Method uses Matrix approach of the form: $A \times X = B$, following Elementary Transformation principle.

3.5 References

(a) S. S. Shastry ”Introductory Methods of Numerical Methods”. (Chp 3)

(b) Steven C. Chapra, Raymond P. Canale ”Numerical Methods for Engineers”.

3.6 Unit End Exercise

Find the solution of the following set of equation by using Gauss Jordan method.

(a) $3x + 2y + 4z = 7$ $2x + y + z = 7$ $x + 3y + 5z = 2$

(b) $10x + y + z = 12$ $2x + 10y + z = 13$ $x + y + 3z = 5$

(c) $4x + 3y - z = 6$ $3x + 5y + 3z = 4$ $x + y + z = 1$

(d) $2x + y - z = -1$ $x - 2y + 3z = 9$

$3x - y + 5z = 14$

Find the solution of the following set of equation by using Gauss Seidel method.

(a) $10x + y + z = 12$ $2x + 10y + z = 13$ $x + y + 3z = 5$

(b) $28x + 4y - z = 32$ $2x + 17y + 4z = 35$ $x + 3y + 10z = 24$

(c) $7x_1 + 2x_2 - 3x_3 = -12$ $2x_1 + 5x_2 - 3x_3 = -20$ $x_1 - x_2 - 6x_3 = -26$

(d) $7x_1 + 2x_2 - 3x_3 = -12$ $2x_1 + 5x_2 - 3x_3 = -20$ $x_1 - x_2 - 6x_3 = -26$

(e) $10x + y + z = 12$ $x + 10y + z = 12$ $x + y + 10z = 12$

(Assume $x^{(0)} = 0.4$, $y^{(0)} = 0.6$, $z^{(0)} = 0.8$)

❖❖❖

UNIT 2

4

NUMERICAL DIFFERENTIATION AND INTEGRATION

Unit Structure

4.0 Objectives

Student will be able to understand the following from the Chapter:

Methods to compute value of Differentiation of a function at a particular value.

Methods to find the area covered by the curve within the the given interval.

Understand the importance of Interpolation Method.

4.1 Introduction

Differentiation is the method used to determine the slope of the curve at a particular point. Whereas, Integration is the method used to find the area between two values.

The solution of Differentiation and Integration at and between particular values can be easily determined by using some Standard rules. There are some complex function whose differentiation and Integration solution is very difficult to determine. Hence, there are some practical approaches which can be used to find the approximated value.

4.2 Numerical Differentiation

The value of differentiation or derivative of a function at a particular value can be determined by using Interpolation Methods.

(a) Newton's Difference Method (If the step size is constant.)

(b) La-grange's Interpolation Method (If step size is not constant.)

Newton's Difference Method is further divided into two types depending on the position of the sample present in the input data set.

Newton's **Forward** Difference Method: To be used when the input value is lying towards the **start** of the input data set.

Newton's **Backward** Difference Method: To be used when the input value is lying at the **end** of the input data set.

4.2.1 *Newton Forward Difference Method*

For a given set of values (x_i, y_i), $i = 0,1,2,...n$ where x_i are at equal intervals and h is the interval of of the input values i.e. $x_i = x_0 + i \times h$ then,

$$y(x) = y_0 + p \triangle y_0 + \frac{p(p-1)}{2!} \triangle^2 y_0 + \frac{p(p-1)(p-2)}{3!} \triangle^3 y_0 + \cdots\cdots + \\ + \frac{p(p-1)(p-2)\cdot(p-n+1)}{n!} \triangle^n y_0$$

Since the above equation is in form of "p", therefore, chain rule is to be used for differentiation. Hence,

$$\frac{dy}{dx} = \frac{dy}{dp} \times \frac{dp}{dx}$$

We know that,

$$p = \frac{x - x_0}{h}$$

Where, x = Unknown Value x_0 = First Input value h = Step Size

Hence,

$$\frac{dp}{dx} = \frac{1}{h}$$

And,

$$\frac{dy}{dp} = \triangle y_0 + \frac{p-1+p}{2!}\triangle^2 y_0 + \frac{p(p-1)+p(p-2)+(p-1)(p-2)}{3!}\triangle^3 y_0 + \cdots\cdots$$

Hence due to simplification,

$$\frac{dy}{dx} = \frac{1}{h} \times \left(\triangle y_0 - \frac{\triangle^2 y_0}{2} + \frac{\triangle^3 y_0}{3} - \frac{\triangle^4 y_0}{4} + \frac{\triangle^5 y_0}{5} - \frac{\triangle^6 y_0}{6} + \cdots \right)$$

On differentiating second time, the expression becomes:

$$\frac{d^2y}{dx^2} = \frac{1}{h^2} \times \left(\triangle^2 y_0 - \triangle^3 y_0 + \frac{11\,\triangle^4 y_0}{12} - \frac{5\,\triangle^5 y_0}{6} + \frac{137\,\triangle^6 y_0}{180} - \cdots \right)$$

4.2.2 *Newton Backward Difference Method*

For a given set of values (x_i,y_i), $i = 0,1,2,...n$ where x_i are at equal intervals and h is the interval of of the input values i.e. $x_i = x_0 + i \times h$ then,

$$y(x) = y_n + p \bigtriangledown y_n + \frac{p(p+1)}{2!} \bigtriangledown^2 y_n + \frac{p(p+1)(p+2)}{3!} \bigtriangledown^3 y_n + \cdots\cdots + \cdots\cdots + \frac{p(p+1)(p+2)\cdot(p+n-1)}{n!} \bigtriangledown^n y_0$$

Since the above equation is in form of "p", therefore, chain rule is to be used for differentiation. Hence,

$$\frac{dy}{dx} = \frac{dy}{dp} \times \frac{dp}{dx}$$

We know that,

$$p = \frac{x - x_n}{h}$$

Where, x = Unknown Value x_n = Final Input value h = Step Size

Hence,

$$\frac{dp}{dx} = \frac{1}{h}$$

And,

$$\frac{dy}{dp} = \nabla y_n + \frac{p+1+p}{2!}\nabla^2 y_n + \frac{p(p+1)+p(p+2)+(p+1)(p+2)}{3!}\nabla^3 y_n + \cdots\cdots$$

Hence due to simplification,

$$\frac{dy}{dx} = \frac{1}{h} \times \left(\nabla y_n + \frac{\nabla^2 y_n}{2} + \frac{\nabla^3 y_n}{3} + \frac{\nabla^4 y_n}{4} + \frac{\nabla^5 y_n}{5} + \frac{\nabla^6 y_n}{6} + \cdots \right)$$

On differentiating second time, the expression becomes:

$$\frac{d^2y}{dx^2} = \frac{1}{h^2} \times \left(\nabla^2 y_n + \nabla^3 y_n + \frac{11\,\nabla^4\, y_n}{12} + \frac{5\,\nabla^5\, y_n}{6} + \frac{137\,\nabla^6\, y_n}{180} + \cdots \right)$$

4.2.3 *Solved Examples*

i. From the data table given below obtain $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ at $x = 1.2$

x	1.0	1.2	1.4	1.6	1.8	2.0	2.2
y	2.7183	3.3201	4.0552	4.9530	6.0496	7.3891	9.0250

Sol. The first step is to identify which method to be used.

Since in the question the value of $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ at $x = 1.2$ is to be determined.

Hence, **Forward Difference** to be used as the value lies at the start of the data set.

Therefore, Forward Difference Table is to be formed.

x	**y**	$4y$	4^2y	4^3y	4^4y	4^5y	4^6y
1	2.7183	0.6018	0.1333	0.0294	0.0067	0.0013	0.0001
1.2	3.3201	0.7351	0.1627	0.0361	0.0080	0.0014	
1.4	4.0552	0.8978	0.1988	0.0441	0.0094		
1.6	4.9530	1.0966	0.2429	0.0535			
1.8	6.0496	1.3395	0.2964				
2.0	7.3891	1.6359					
2.2	9.0250						

Now select the values corresponding to the row of $x = 1.2$. Since, value of $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ is to be determined at $x = 1.2$

Hence, the values are: $4y_0 = 0.7351$, $4^2y_0 = 0.1627$, $4^3y_0 = 0.0361$, $4^4y_0 = 0.0080$, $4^5y_0 = 0.0014$.

Substituting the values in the following formula:

$$\frac{dy}{dx} = \frac{1}{h} \times \left(\triangle y_0 - \frac{\triangle^2 y_0}{2} + \frac{\triangle^3 y_0}{3} - \frac{\triangle^4 y_0}{4} + \frac{\triangle^5 y_0}{5} \right)$$

$$\frac{dy}{dx} = 3.2031$$

Similarly, to find the value of $\frac{d^2y}{dx^2}$, Substitute the values in the following equation:

$$\frac{d^2y}{dx^2} = \frac{1}{h^2} \times \left(\triangle^2 y_0 - \triangle^3 y_0 + \frac{11 \triangle^4 y_0}{12} - \frac{5 \triangle^5 y_0}{6} \right)$$

where, $\triangle^2 y_0 = 0.1627$, $\triangle^3 y_0 = 0.0361$, $\triangle^4 y_0 = 0.0080$, $\triangle^5 y_0 = 0.0014$.

Therefore,

$$\frac{d^2y}{dx^2} = 3.3191$$

i. From the data table given below obtain $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ at $x = 1.8$

x	1.0	1.2	1.4	1.6	1.8	2.0	2.2
y	2.7183	3.3201	4.0552	4.9530	6.0496	7.3891	9.0250

Sol. The first step is to identify which method to be used.

Since in the question the value of $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ at $x = 1.8$ is to be determined.

Hence, **Backward Difference** to be used as the value lies at the end of the data set.

Therefore, Backward Difference Table is to be formed.

x	**y**	$5y$	5^2y	5^3y	5^4y	5^5y	5^6y
1	2.7183						
1.2	3.3201	0.6018					
1.4	4.0552	0.7351	0.1333				

x	y	$5y$	5^2y	5^3y	5^4y	5^5y	5^6y
1.6	4.9530	0.8978	0.1627	0.0294			
1.8	6.0496	1.0966	0.1988	0.0067			
2.0	7.3891	1.3395	0.2429	0.0441	0.0080	0.0013	
2.2	9.0250	1.6359	0.2964	0.0535	0.0094	0.0014	0.0001

Now select the values corresponding to the row of $x = 1.2$. Since,

value of $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ is to be determined at $x = 1.8$

Hence, the values are: $\nabla y_n = 1.0966,\ \nabla^2 y_n = 0.1988,\ \nabla^3 y_n = 0.0361,\ \nabla^4 y_n = 0.0067.$

Substituting the values in the following formula:

$$\frac{dy}{dx} = \frac{1}{h} \times \left(\nabla y_n + \frac{\nabla^2 y_n}{2} + \frac{\nabla^3 y_n}{3} + \frac{\nabla^4 y_n}{4} \right)$$

$\frac{dy}{dx} = 6.04854.$

Similarly, to find the value of $\frac{d^2y}{dx^2}$, Substitute the values in the following equation:

$$\frac{d^2y}{dx^2} = \frac{1}{h^2} \times \left(\nabla^2 y_n + \nabla^3 y_n + \frac{11\ \nabla^4\ y_n}{12} \right)$$

where, $\nabla^2 y_n = 0.1988,\ \nabla^3 y_n = 0.0361,\ \nabla^4 y_n = 0.0067.$

Therefore,

$\frac{d^2y}{dx^2} = 3.3191$

If **step size** of the input value is **not constant**, then the derivatives can be determined by simply differentiating the expression provided by the **Langrange's Polynomial**.

4.2.4 *Solved Examples*

i. Tabulate the following function: $y = x^3 - 10x + 6$ at $x_0 = 0.5$, $x_1 = 1$ and $x_2 = 2$. Compute its 1^{st} and 2^{nd} derivatives at $x = 1.00$ using Lagrange's interpolation method.

Sol. Given: $y = x^3 - 10x + 6$

At $x_0 = 0.5$ $y_0 = 0.5^3 - 10 \times 0.5 + 6 = 1.125$

At $x_1 = 1$ $y_1 = 1^3 - 10 \times 1 + 6 = -3$

At $x_2 = 2$ $y_2 = 2^3 - 10 \times 2 + 6 = -6$

Hence, the Lagange's Formula is:

$$y = \frac{(x-1)(x-2)}{(0.5-1)(0.5-2)} \times (1.125) + \frac{(x-0.5)(x-1)}{(2-0.5)(2-1)} \times (-6) + \frac{(x-0.5)(x-2)}{(1-0.5)(1-2)} \times (-3)$$

$$y = \frac{(x-1)(x-2)}{0.75} \times (1.125) + \frac{(x-0.5)(x-1)}{1.5} \times (-6) + \frac{(x-0.5)(x-2)}{0.5} \times (3)$$

Hence, $\frac{dy}{dx}$ will be obtained by differentiating both the sides

$$\frac{dy}{dx} = \frac{(x-1)+(x-2)}{0.75} \times (1.125) + \frac{(x-0.5)+(x-1)}{1.5} \times (-6) + \frac{(x-0.5)+(x-2)}{0.5} \times (3)$$

So therefore, at $x = 1$ the $\frac{dy}{dx}$ will be:

$$\frac{dy}{dx} = \frac{(1-1)+(1-2)}{0.75} \times (1.125) + \frac{(1-0.5)+(1-1)}{1.5} \times (-6) + \frac{(1-0.5)+(1-2)}{0.5} \times (3)$$

$$\frac{dy}{dx} = \frac{-1.125}{0.75} + \frac{-0.5}{1.5} \times (6) + (-3)$$

$$\frac{dy}{dx} = -6.5$$

4.3 Numerical Integration

Numerical Integration provides a set of methods to compute the Definite Integration of a Function between the given set of values.

There are three methods used to find the value of Integration:

- Trapezoidal Rule.
- Simpson's $\frac{1}{3^{rd}}$ Rule.
- Simpson's $\frac{3}{8^{th}}$ Rule.

4.3.1 Trapezoidal Rule

In this method the curve is divided into small trapeziums.

These trapeziums are then added to to find the complete area of the curve between two values. Hence,

Let $y = f(x)$ then,

$$\int_{x_0}^{x_n} f(x)dx = \frac{h}{2}\Big(y_0 + y_n + 2 \times (y_1 + y_2 + \ldots + y_{n-1})\Big)$$

Where,

$$h = \frac{x_n - x_0}{n}$$

x_n: Upper Limit x_0: Lower Limit $y_0, y_1, y_2, \cdots, y_n$ are the values of of y corresponding to $x_0, x_1, x_2, \cdots, x_n$

4.3.2 Simpson's $\frac{1}{3^{rd}}$ Rule

Let $y = f(x)$ then,

$$\int_{x_0}^{x_n} f(x)dx = \frac{h}{3}\Big(y_0 + y_n + 4 \times \text{(Sum of Odd osition Terms)} + 2 \times$$

(Sum of Even osition Terms)

$$\int_{x_0}^{x_n} f(x)dx = \frac{h}{3}\Big(y_0 + y_n + 4 \times (y_1 + y_3 + y_5 + \cdots) + 2 \times (y_2 + y_4 + y_6 + \cdots)\Big)$$

Where,

$$h = \frac{x_n - x_0}{n}$$

x_n: Upper Limit

x_0: Lower Limit

$y_0, y_1, y_2, \cdots, y_n$ are the values of of y corresponding to $x_0, x_1, x_2, \cdots, x_n$

4.3.2 Simpson's $\frac{3}{8^{th}}$ Rule

Let $y = f(x)$ then,

$$\int_{x_0}^{x_n} f(x)dx = \frac{3h}{8}\Big(y_0 + y_n + 2 \times \text{(Sum of Multiple of 3, position terms)} +$$

3 × (Sum of Remaining terms)

$$\int_{x_0}^{x_n} f(x)dx =$$

$$\frac{3h}{8}\Big(y_0 + y_n + 2 \times (y_3 + y_6 + y_9 + \cdots) + 3 \times (y_1 + y_2 + y_4 + y_5 + \cdots)\Big)$$

Where,

$$h = \frac{x_n - x_0}{n}$$

x_n: Upper Limit

x_0: Lower Limit

$y_0, y_1, y_2, \cdots, y_n$ are the values of of y corresponding to $x_0, x_1, x_2, \cdots, x_n$

4.3.3 *Solved Examples*

i. A solid of revolution is formed by rotating about the x-axis, the area between the x-axis,the lines $x = 0$ and $x = 1$ and the curve through the points below:

x	0.00	0.25	0.50	0.75	1.00
x	1.000	0.9896	0.9589	0.9089	0.8415

Estimate the volume of the solid formed.

Sol. Let "V" be the volume of the solid formed by rotating the curve around x- axis, then

$$V = \pi \times \int_0^1 y^2 dx$$

Therefore, the tables is updated as:

x	0.00	0.25	0.50	0.75	1.00
x	1.000	0.9793	0.9195	0.8261	0.7081

Rewriting the same table to find the value of n.

x	$0.00 = x_0$	$0.25 = x_1$	$0.50 = x_2$	$0.75 = x_3$	$1.00 = x_4$
x	1.000	0.9793	0.9195	0.8261	0.7081

As the extreme or the last value is x_4. Hence, $n = 4$

Therefore,

$$h = \frac{1 - 0}{4}$$
$$h = \frac{1}{4}$$
$$h = 0.25$$

(a) TRAPEZOIDAL RULE

$$\pi \times \int_0^1 y^2 dx = \pi \times \frac{h}{2}\Big(y_0 + y_4 + 2 \times (y_1 + y_2 + y_3)\Big)$$

where, $h = 0.25$, $y_0 = 1.000$, $y_1 = 0.9793$, $y_2 = 0.9195$, $y_3 = 0.8261$, $y_4 = 0.7081$

On substituting the values we get:

$$\pi \times \int_0^1 y^2 dx = 3.36704$$

(b) SIMPSON'S $\frac{1}{3^{rd}}$ RULE

$$\pi \times \int_0^1 y^2 dx = \pi \times \frac{h}{3}\left(y_0 + y_4 + 4 \times (y_1 + y_3) + 2 \times (y_2)\right)$$

where, $h = 0.25$, $y_0 = 1.000$, $y_1 = 0.9793$, $y_2 = 0.9195$, $y_3 = 0.8261$, $y_4 = 0.7081$

On substituting the values we get:

$$\pi \times \int_0^1 y^2 dx = 2.81923$$

(c) SIMPSON'S $\frac{3}{8^{th}}$ RULE

$$\pi \times \int_0^1 y^2 dx = \pi \times \frac{3h}{8}\left(y_0 + y_4 + 2 \times (y_3) + 3 \times (y_1 + y_2)\right)$$

where, $h = 0.25, y_0 = 1.000, y_1 = 0.9793, y_2 = 0.9195, y_3 = 0.8261, y_4 = 0.7081$

On substituting the values we get:

$$\pi \times \int_0^1 y^2 dx = 2.66741$$

4.4 Summary

Numerical Differentiation uses the methods to find the value of First ans Second Order Derivatives at a particular value of x or the input variable.

Numerical Integration provides methods to find the Definite Integration or the area covered by the curve between two points.

4.5 References

(a) S. S. Shastry "Introductory Methods of Numerical Methods".

(b) Steven C. Chapra, Raymond P. Canale "Numerical Methods for Engineers".

4.6 Unit End Exercise

(a) Find $\frac{d(J_0(x))}{dx}$ and $\frac{d^2(J_0(x))}{dx^2}$ at $x = 0.1$, $x = 0.2$ and $x = 0.4$ from the following table:

(0,1.0); (0.1,0.9975); (0.2,0.9900); (0.3,0.9776); (0.4,0.9604)

(b) The following table gives angular displacement a at different time t (time).
(0,0.052); (0.02,0.105); (0.04,0.168); (0.06,0.242); (0.08,0.327) ;
(0.10,0.408) ;

Calculate angular velocity and acceleration at t=0.04, 0.06, and 0.1

Angular Velocity :$\frac{d\theta}{dt}$; Angular Acceleration :$\frac{d^2\theta}{dt^2}$

(c) A cubic function $y = f(x)$ satisfies the following data.

x	0	1	3	4
y	1	4	40	85

Determine $f(x)$ and hence find $f'(2)$ and $f''(2)$

(d) Use Trapezoidal Rule to evaluate integral $\int_0^2 e^{-x}dx$ width of sub-interval (h) = 0.5.

(e) Using Simpson's rules. Evaluate $\int_0^6 \left(\frac{1}{x^4+1}\right) dx$ take $n = 6$.

(f) Using Simpson's rules. Evaluate $\int_{-3}^3 (x^4)\, dx$ take $n = 6$.

(g) Using Simpson's rules. Evaluate $\int_0^6 \left(\frac{1}{x+1}\right) dx$ take $n = 6$.

(h) Use Trapezoidal Rule to evaluate integral $\int_0^2 x \times e^{-x}dx$ width of sub-interval (h) = 0.5.

❖❖❖

UNIT 2

5

NUMERICAL DIFFERENTIATION EQUATION

Unit Structure

5.0 Objectives

Student will be able to understand the following from the Chapter:

Methods to compute value of Differential Equation of a function at a particular value.

Practical or Software Implemented method to find the solution of Differential equation.

5.1 Introduction

Differential equation is defined as an expression which contains derivative terms. For example: $\frac{dy}{dx} = 2x + 3y$. Here the term $\frac{dy}{dx}$ is indicating that the occurrence of a Differential Equation.

The differential equations can be solved analytically using various methods like: *Variable separable*, *Substitution*, *Linear Differential Equation*, *Solution to Homogeneous Equation*, etc. but in this chapter, various practically approachable methods will be discussed to find the solution of a given differential equation at a particular value.

A Differential equation is defined on the basic of two terminologies:

Order: Number of times a variable is getting differentiated by an another variable.

Degree: Power of the highest order derivative in a differential equation.

For Example: Consider the following equations: A. $\frac{dy}{dx} = 4x + 5y$

B. $\frac{d^2y}{dx^2} = 5x^2 + 6xy$ C. $\left(\frac{dy}{dx}\right)^2 = 5x^3 + 6y^2$

D. $\frac{d^2y}{dx^2} + \left(\frac{dy}{dx}\right)^2 = x - y$

In the above equations, eqn. *A* has **Order** = 1, because *y* is differentiated only 1 time w.r.t. *x*. The **degree** = 1 because the power of the only differential equation is 1.

In eqn. *B* has **Order** = 2, because *y* is differentiated 2 times w.r.t. *x*. The **degree** = 1 because the power of the only differential equation is 1.

In eqn. *C* has **Order** = 1, because *y* is differentiated only 1 time w.r.t. *x*. The **degree** = 2 because the power of the only differential equation is 2.

In eqn. *D*, there are two derivatives present. Hence, the term with maximum differentiation present is to be selected. Therefore, the equation has **Order** = 2, because *y* is differentiated only 2 time w.r.t. *x*. The **degree** = 1 because the power of the derivative of maximum order is 1.

5.2 Euler's Method

Euler's Method is a Practically implemented method used to find the solution of **first order** Differential equation.

Suppose the given differential equation is $\frac{dy}{dx} = f(x, y)$ with initial conditions $y(x_0) = y_0$ (The value of y at $x = x_0$ is y_0.) Then by Euler's method:

$$y_{n+1} = y_n + h \times f(x_n, y_n) \text{ for } n = 0, 1, 2, \ldots$$

Where,

y_{n+1} is *Future Value of y*. y_n is *Present Value of y*.

h is *Common Difference or Step Size*.

5.2.1 Solved Examples

Find the value of *y* when $x = 0.1$, given that $y(0) = 1$ and $y^0 = x^2 + y$ by using Euler's Method.

Sol. Given: $y^0 = x^2 + y$ with initial condition $y(0) = 1$

$$y' = \frac{dy}{dx} = x^2 + y$$ and the meaning of $y(0) = 1$ is the value of *y* at $x = 0$

is 1. Hence,

$x_0 = 0$ and $y_0 = 1$

To find the value of *y* at $x = 0.1$, value of *h* is required.

Hence, $h = 0.05$.

According to Euler's Method, $y_{n+1} = y_n + h \times f(x_n, y_n)$.

Iteration 1:

$y_1 = y_0 + h \times f(x_0, y_0)$ $y_0 = 1$; $x_0 = 0$; $h = 0.05$ $\therefore f(x_n, y_n) = x^2 + y$ $y_1 = 1 + 0.05(0^2 + 1)$

$y_1 = 1.05$ (The value of *y* at $x = x_0 + h = 0 + 0.05 = 0.05$

Iteration 2:

$y_2 = y_1 + h \times f(x_1, y_1)$ $y_2 = 1.05$; $x_1 = x_0 + h = 0.05$ $y_2 = 1.05 + 0.05(0.05^2 + 1.05)$ $y_1 = 1.0265$ (The value of *y* at $x = x_1 + h = 0.05 + 0.05 = 0.10$

5.3 Euler's Modified Method

The values of *y* determined at every iteration may have some error depending on the value of selected Common Difference or Step Size *h*. Hence, to find the accurate value of *y* at a particular *x*, **Euler's Modified Method** is used.

In this method the value of the corresponding iteration is ensured by providing a process called *Iteration within Iteration method*.

This method is used to minimize the error. The iterative formula is given as;

$$y_{n+1}^{(m+1)} = y_n + \frac{h}{2}\left[f(x_m, y_m) + f\ \ x_{n+1}, y_{n+1}^m \right]$$

Where,

x_m, y_m are the values to be used in the basic iteration and $y(m+1)^n$ is the value obtained while saturating the given "y" value in the same iteration.

The steps to use Euler's Modified Method is:

(a) Apply Euler's Method at every iteration to find the approximate value at new value of x using **Euler's Method**.

(b) Apply **Euler's Modified Method** in a particular iteration to saturate the value of y.

5.3.1 Solved Examples

Find the value of y when $x = 0.1$, given that $y(0) = 1$ and $y^0 = x^2 + y$ by using Euler's Modified Method.

Sol. Given: $y^0 = x^2 + y$ with initial condition $y(0) = 1$

$y' = \frac{dy}{dx} = x^2 + y$ and the meaning of $y(0) = 1$ is the value of y at $x = 0$

is 1. Hence,

$x_0 = 0$ and $y_0 = 1$

To find the value of y at $x = 0.1$, value of h is required.

Hence, $h = 0.05$.

Iteration 1:

Using Euler's Method $y_1 = y_0 + h \times f(x_0,y_0)$ $y_0 = 1$; $x_0 = 0$; $h = 0.05$ $\therefore f(x_n,y_n) = x^2 + y$ $y_1 = 1 + 0.05(0^2 + 1)$ $y_1 = 1.05$

Iteration 1 a :

Using Euler's Modified Method,

$$y_1^1 = y_0 + \frac{h}{2}\left[f(x_0, y_0) + f(x_1, y_1^0)\right]$$

where, $x_1 = x_0 + h = 0 + 0.05 = 0.05$ and y_1^0 is the initial value obtained by using Euler's Method.

$$\therefore y_1^1 = 1 + \frac{0.05}{2}\left[f(0^2 + 1) + f(0.05^2 + 1.05)\right]$$

$$y_1^1 = 1.0513$$

Since, $y_1^0 = 1.0500$ and $y_1^1 = 1.0513$ are not equal. Hence, apply 2^{nd} iteration.

Iteration 1 b :

Using Euler's Modified Method,

$$y_1^2 = y_0 + \frac{h}{2}\left[f(x_0, y_0) + f(x_1, y_1^1)\right]$$

where, $x_1 = x_0 + h = 0 + 0.05 = 0.05$ and $y_1^1 = 1.0513$ is the value obtained in *Iteration 1(a)*.

$$\therefore y_1^2 = 1 + \frac{0.05}{2}\left[f(0^2 + 1) + f(0.05^2 + 1.0513)\right]$$
$$y_1^2 = 1.0513$$

Since, values of y_1^2 and y_1^1 are equal. Hence, the value of y at $x = 0.05$ is 1.0513

Iteration 2:

Using Euler's Method $y_2 = y_1 + h \times f(x_1, y_1)$ $y_1 = 1.0513$; $x_1 = x_0 + h = 0.05$; $h = 0.05$

$\therefore f(x_n, y_n) = x^2 + y$

$y_2 = 1.0513 + 0.05(0.05^2 + 1.0513)$ $y_2 = 1.1044$

Iteration 2 a :

Using Euler's Modified Method,

$$y_2^1 = y_1 + \frac{h}{2}\left[f(x_1, y_1) + f(x_2, y_2^0)\right]$$

where, $x_2 = x_1 + h = 0.05 + 0.05 = 0.10$ and y_2^0 is the initial value obtained by using Euler's Method.

$$\therefore y_2^1 = 1.1044 + \frac{0.05}{2}\left[f(0.05^2 + 1.0513) + f(0.1^2 + 1.1044)\right]$$
$$y_2^1 = 1.1055$$

Since, $y_2^0 = 1.1044$ and $y_2^1 = 1.1055$ are not equal. Hence, apply 2^{nd} iteration.

Iteration 1 b :

Using Euler's Modified Method,

$$y_2^2 = y_1 + \frac{h}{2}\left[f(x_1, y_1) + f(x_2, y_2^1)\right]$$

where, $x_2 = x_1 + h = 0.05 + 0.05 = 0.10$ and y_2^1 is the value obtained at

Iteration 2 a.

$$\therefore y_2^2 = 1.1044 + \frac{0.05}{2}\left[f\left(0.05^2 + 1.1055\right) + f\left(0.1^2 + 1.1055\right)\right]$$
$$y_2^1 = 1.1055$$

Since, values of y_2^2 and y_2^1 are equal. Hence,

the value of y at $x = 0.1$ is 1.1055

5.4 Range Kutta Method

Range Kutta Method is an another method to find the solution of a First Order Differential Equations. It is mainly divided into two methods depending on the number of parameters used in a method.

Range Kutta 2nd Order Method.

Range Kutta 4th Order Method.

Range Kutta 2nd order Method uses **Two** parameters k_1 and k_2 to find the value of y_{n+1}. The expression is given as:

$$y_{n+1} = y_n + \frac{1}{2} \times [k_1 + k_2]$$

where

$k_1 = h \times f(x_n, y_n)$ $k_2 = h \times f(x_n + h, y_n + k_1)$

Range Kutta 4th order Method uses **Four** parameters k_1, k_2, k_3 and k_4 to find the value of y_{n+1}. The expression is given as:

$$y_{n+1} = y_n + \frac{1}{6} \times [k_1 + 2 \times k_2 + 2 \times k_3 + k_4]$$

where,

$$k_1 = h \times f(x_n, y_n)$$
$$k_2 = h \times f\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}\right)$$
$$k_3 = h \times f\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}\right)$$
$$k_2 = h \times f(x_n + h, y_n + k_3)$$

5.4.1 Solved Examples

Given; $\frac{dy}{dx} = y - x$, where $y_0 = 2$, Find $y(0.1)$ and $y(0.2)$, correct upto 4 decimal places.

Sol. Given: $\frac{dy}{dx} = y - x = f(x, y) y(0) = 2$

$\therefore y_0 = 2$ and $x_0 = 0$

Range Kutta 2^{nd} Order Method:

$h = 0.1$

Iteration 1:

$$y_1 = y_0 + \frac{1}{2}(k_1 + k_2)$$
$$k_1 = h \times f(x_0, y_0)$$
$$\therefore k_1 = 0.1(2 - 0)$$
$$\therefore k_1 = 0.2$$
$$k_2 = h \times f(x_0 + h, y_0 + k_1)$$
$$\therefore k_2 = 0.1 \times ((2 + 0.2) - (0 + 0.1))$$
$$\therefore k_2 = 0.21$$

So, $y_1 = 2 + \frac{1}{2}(0.2 + 0.21)$

$\therefore y_1 = 2.2050$ at $x = 0.1$ **Iteration 2:**

$$y_2 = y_1 + \frac{1}{2}(k_1 + k_2)$$

$k_1 = h \times f(x_1, y_1)$

$\therefore k_1 = 0.1(2.205 - 0.1)$ $\therefore k_1 = 0.2105$ $k_2 = h \times f(x_1 + h, y_1 + k_1)$

$\therefore k_2 = 0.1 \times ((2.2050 + 0.2105) - (0.1 + 0.1))$

$\therefore k_2 = 0.22155$

So, $y_1 = 2.205 + \frac{1}{2}(0.2105 + 0.22155)$

$\therefore y_1 = 2.421025$ at $x = 0.2$ **Runge Kutta 4^{th} Order:**

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2 \times k_2 + 2 \times k_3 + k_4)$$

Where,

$$k_1 = h \times f(x_n, y_n)$$
$$k_2 = h \times f\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}\right)$$
$$k_3 = h \times f\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}\right)$$
$$k_2 = h \times f(x_n + h, y_n + k_3)$$

Let h=0.1,

Iteration 1:

$$y_1 = y_0 + \frac{1}{6}[k_1 + 2 \times k_2 + 2 \times k_3 + k_4]$$
$$k_1 = h \times f(x_0, y_0)$$
$$\therefore k_1 = h \times f(x_0, y_0)$$
$$\therefore k_1 = 0.1(2 - 0)$$
$$\therefore k_1 = 0.2$$
$$k_2 = h \times f\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right)$$
$$\therefore k_2 = 0.1 \times \left(\left(2 + \frac{0.2}{2}\right) - \left(0 + \frac{0.1}{2}\right)\right)$$
$$\therefore k_2 = 0.215$$
$$k_3 = h \times f\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right)$$
$$\therefore k_3 = 0.1 \times \left(\left(2 + \frac{0.215}{2}\right) - \left(0 + \frac{0.1}{2}\right)\right)$$
$$\therefore k_3 = 0.20575$$

$$k_4 = h \times f(x_0 + h, y_0 + k_3)$$
$$\therefore k_4 = 0.1 \times ((2 + 0.20575) - (0 + 0.1))$$
$$\therefore k_4 = 0.210575$$
$$\therefore y_1 = y_0 + \frac{1}{6}[k_1 + 2 \times k_2 + 2 \times k_3 + k_4]$$
$$\therefore y_1 = 2 + \frac{1}{6}[0.2 + 2 \times 0.215 + 2 \times 0.20575 + 0.210575]$$
$$\therefore y_1 = 2.2087 \text{ at } x_1 = 0.1$$

Iteration 2:

$$y_2 = y_1 + \frac{1}{6}\left[k_1 + 2 \times k_2 + 2 \times k_3 + k_4\right]$$
$$k_1 = h \times f\left(x_1, y_1\right)$$
$$\therefore k_1 = h \times f(x_1, y_1)$$
$$\therefore k_1 = 0.1(2.2087 - 0.1)$$
$$\therefore k_1 = 0.21087$$
$$k_2 = h \times f\left(x_1 + \frac{h}{2}, y_1 + \frac{k_1}{2}\right)$$
$$\therefore k_2 = 0.1 \times \left(\left(2.2087 + \frac{0.21087}{2}\right) - \left(0.1 + \frac{0.1}{2}\right)\right)$$
$$\therefore k_2 = 0.2164$$
$$k_3 = h \times f\left(x_1 + \frac{h}{2}, y_1 + \frac{k_2}{2}\right)$$
$$\therefore k_3 = 0.1 \times \left(\left(2.2087 + \frac{0.2164}{2}\right) - \left(0.1 + \frac{0.1}{2}\right)\right)$$
$$\therefore k_3 = 0.21669$$
$$k_4 = h \times f\left(x_1 + h, y_1 + k_3\right)$$
$$\therefore k_4 = 0.1 \times ((2.2087 + 0.21669) - (0.1 + 0.1))$$
$$\therefore k_2 = 0.222539$$
$$\therefore y_2 = y_1 + \frac{1}{6}\left[k_1 + 2 \times k_2 + 2 \times k_3 + k_4\right]$$
$$\therefore y_1 = 2.2087 + \frac{1}{6}\left[0.21087 + 2 \times 0.2164 + 2 \times 0.21669 + 0.222539\right]$$
$$\therefore y_1 = 2.4253$$

5.5 Taylor Series

The methods discussed in the previous sections, are applicable only for **First** order Differential Equation. But, Taylor Series method can be used to find the solution for higher order differential equations.

Taylor Series is a method used to represent a Function as a sum of infinite series represented in terms of derivatives derived at a particular point. Taylor series is mathematically expressed as:

$$f(x) = f(x_0) + \frac{xf'(x_0)}{1!} + \frac{x^2 f''(x_0)}{2!} + \frac{x^3 f'''(x_0)}{3!} \cdots$$

Similarly, to get an approximate value of y at $x = x_0 + h$ is given by following expression:

$$y(x_0 + h) = y_0 + hy_0' + \frac{h^2 y_0''}{2!} + \frac{h^3 y_0'''}{3!} + \cdots$$

5.5.1 Solved Examples

Example 12. Given the differential equation $y'' - xy' - y = 0$, with the condition $y(0) = 1$ and $y'(0) = 0$, Use Taylor series to determine the value of $y(0.1)$.

Solution: $y'' - xy' - y = 0$

Now, $y(0) = 1$ and $y'(0) = 0$ and also $x_0 = 0$

According to Taylor Series;

$$y(x) = y_0 + (x - x_0)\, y_0' + \frac{(x - x_0)^2}{2!} \times y_0'' + \frac{(x - x_0)^3}{3!} \times y_0''' + \frac{(x - x_0)^4}{4!} \times y_0^{iv} + \frac{(x - x_0)^5}{5!} \times y_0^{v}$$

Since, $y'' - xy' - y = 0$

$\therefore y'' = xy' + y \ldots\ldots (1)$

$\therefore y''(0) = x_0 y_0' + y_0$

$\therefore y''(0) = 0 \times 0 + 1$

$\therefore y''(0) = 1$

Differentiating Equation (1)

$y''' = y' + xy'' + y' \ldots\ldots (2)$

$\therefore y'''(0) = y_0' + x_0 y_0'' + y_0'$

$\therefore y'''(0) = 0 + 0 \times 1 + 0$

$\therefore y'''(0) = 0$

Differentiating Equation (2)

$y^{iv} = y'' + y'' + xy''' + y''$

$y^{iv} = 3y'' + xy''' \ldots\ldots (3)$

$\therefore y^{iv}(0) = 3y_0'' + x_0 y_0'''$

$\therefore y^{iv}(0) = 3 \times 1 + 0 \times 1 + 0$

$\therefore y^{iv}(0) = 3$

Differentiating Equation (3)

$y^{v} = 3y''' + y''' + xy^{iv}$

$y^{v} = 4y''' + xy^{iv} \ldots\ldots (4)$

$\therefore y^{v}(0) = 4y_0''' + x_0 y_0^{iv}$

$\therefore y^{v}(0) = 4 \times 0 + 0 \times 3$

$\therefore y^{v}(0) = 0$

Since, $y(x) = y_0 + (x)\, y_0' + \frac{(x)^2}{2!} \times y_0'' + \frac{(x)^3}{3!} \times y_0''' + \frac{(x)^4}{4!} \times y_0^{iv} + \frac{(x)^5}{5!} \times y_0^{v}$

Because $x_0 = 0$

Where, $y_0 = 1, y_0' = 0$, $y_0'' = 1$, $y_0''' = 0$, $y_0^{iv} = 3$ and $y_0^{v} = 0$

$$\therefore y(0.1) = 1 + 0.1 \times 0 + \frac{(0.1)^2}{2!} \times 1 + \frac{(0.1)^3}{3!} \times 0 + \frac{(0.1)^4}{4!} \times 3 + \frac{(0.1)^5}{5!} \times 0$$

$$y(0.1) = 1 + \frac{(0.1)^2}{2!} \times 1 + \frac{(0.1)^4}{4!} \times 3$$

$\therefore y(0.1) = 1.00501$

5.6 Summary

Euler's Method, Euler's Modified and Range Kutta Method are applicable only in First Order Differential Equation.

The First order differential equation should be of the form $\frac{dy}{dx} = f(x, y)$

Solution of Higher Order Differential Equation can be done by using Taylor's Method.

5.7 References

(a) S. S. Shastry "Introductory Methods of Numerical Methods".

(b) Steven C. Chapra, Raymond P. Canale "Numerical Methods for Engineers".

5.7 Unit End Exercise

(a) Use Euler's method to estimate $y(0.5)$ of the following equation with $h = 0.25$ and $\frac{dy}{dx} = x + y + xy,\ y(0) = 1.$

(b) Apply Euler's method to solve $\frac{dy}{dx} = x + 3y$ with $y(0) = 1$. Hence, find $y(1)$. Take $h = 0.2$

(c) Using Euler's method, find $y(2)$ where $\frac{dy}{dx} = 2y + x$ and $y(1) = 1$, take $h = 0.2$.

(d) Using Euler's method, find $y(2)$ where $\frac{dy}{dx} = 2 + \sqrt{xy}$ and $y(1) = 1$, take $h = 0.2$.

(e) Solve $y^0 = 1 - y$, $y(0) = 0$ by Euler's Modified Method and obtain y at $x = 0.1$ and $x = 0.2$

(f) Apply Euler's Modified method to find $y(1.2)$. Given $\frac{dy}{dx} = \frac{y + xy}{x},\ h =$ 0.1 and $y(1) = 2.718$. Correct upto three decimal places.

(g) Solve $\frac{dy}{dx} = ln(x + y),\ y(1) = 2$. Compute y for $x = 1.2$ and $x = 1.4$ using

Euler's Modified Method.

(h) Using Range-Kutta Method of second order to find $y(0.2)$. Given $\frac{dy}{dx} = x + y$, $y(0) = 2$, $h = 0.1$

(i) Use Range-Kutta Method of fourth order to find $y(0.1)$, $y(0.2)$. Given $\frac{dy}{dx} = \frac{y}{2} - 3x$, $y(0) = 1$. (Take $h = 0.1$)

(j) Use Range-Kutta Method of fourth order to find $y(0.1)$. Given $\frac{dy}{dx} = \frac{y^2 + x^2}{10}$, $y(0) = 1$. (Take $h = 0.1$)

(k) Use Range-Kutta Method of second order to find $y(1.2)$. Given $\frac{dy}{dx} = y^2 + x^2$, $y(1) = 0$. (Take $h = 0.1$)

(l) Solve $\frac{dy}{dx} = \frac{y - x}{y + x}$, where $y(0) = 1$, to find $y(0.1)$ using Range-Kutta method.

(m) By using Runge-Kutta method of order 4 to evaluate $y(2.4)$ from the following differential equation $y^0 = f(x,y)$ where $f(x,y) = (x+1)y$. Initial condition $y(2) = 1$, $h = 0.2$ correct upto 4 decimal places.

(n) Use Taylor series method, for the equation, $\frac{dy}{dx} = y - xy$ and $y(0) = 2$ to find the value of y at $x = 1$.

(o) Use Taylor series method, for the equation, $\frac{dy}{dx} = xy + y^2$ and $y(0) = 1$ to find the value of y at $x = 0.1, 0.2, 0.3$.

(p) Use Taylor series method, for the equation, $\frac{dy}{dx} = \frac{1}{y + x^2}$ and $y(4) = 4$ to find the value of y at $x = 4.1, 4.2$.

(q) Use Taylor's series method, for the equation, $y^0 = x^2 - y$ and $y(0) = 1$ to find $y(0.1)$

❖❖❖

UNIT 3

6

RANDOM VARIABLES

Unit Structure

6.0 Objectives

After going through this unit, you will be able to:

- Understand the concept of random variables as a function from sample space to real line
- Understand the concept of probability distribution function of a discrete random variable.
- Calculate the probabilities for a discrete random variable.
- Understand the Probability Distribution of Random Variable
- Understand the probability mass function and probability density function
- Understand the properties of random variables.
- Make familiar with the concept of random variable
- Understand the concept of cumulative distribution function.
- Understand the expected value and variance of a random variable with it s properties

Andrey Nikolaevich Kolmogorov (Russian: 25 April 1903 – 20 October 1987) who advances many scientific fields. In probability theory and related fields, a stochastic or random process is a mathematical object usually defined as a family of random variables. A stochastic process may involve several related random variables. A quotation attributed to Kolmogorov is " Every mathematician believes that he is ahead over all others . The reason why they don't say this in public, is because they are intelligent people."

6.1 Introduction

Many problems are concerned with values associated with outcomes of random experiments. For example we select five items randomly from the basket with

known proportion of the defectives and put them in a packet. Now we want to know the probability that the packet contain more than one defective. The study of such problems requires a concept of a random variable. Random variable is a function that associates numerical values to the outcomes of experiments.

In this chapter, we considered discrete random variable that is random variable which could take either finite or countably infinite values. When a random variable X is discrete, we can assign a positive probability to each value that X can take and determine the probability distribution for X. The sum of all the probabilities associate with the different values of X is one. However not all experiments result in random variables that ate discrete. There also exist random variables such as height, weights, length of life an electric component, time that a bus arrives at a specified stop or experimental laboratory error. Such random variables can assume infinitely many values in some interval on the real line. Such variables are called continuous random variables. If we try to assign a positive probability to each of these uncountable values, the probabilities will no longer sum to one as was the case with discrete random variable.

6.2 Random Variable (R.V.) – Discrete and Continuous

6.2.1 Discrete Random Variable:

Sample space S or Ω contains non-numeric elements. For example, in an experiment of tossing a coin, Ω = { H, T } . However, in practice it is easier to deal with numerical outcomes. In turn, we associate real number with each outcome. For instance, we may call H as 1 and T as 0. Whenever we do this, we are dealing with a function whose domain is the sample space Ω and whose range is the set of real numbers . Such a function is called a random variable.

Definition 1 : Random variable : Let S / Ω be the sample space corresponding to the outcomes of a random experiment . A function X: S → R (where R is a set of real numbers) is called as a random variable.

Random variable is a real valued mapping. A function can either be one-to-one or many-to-one correspondence. A random variable assigns a real number to each possible outcome of an experiment. A random variable is a function from the sample space of a random experiment (domain) to the set of real numbers (co-domain).

Note: Random variable are denoted by capital letters X,Y, Z etc.., where as the values taken by them are denoted by corresponding small letters x, y, z etc.

Definition 2: Discrete random variable : A random variable X is said to be discrete if it takes finite or countably infinite number of possible values. Thus discrete random variable takes only isolated values.

Example1:

What are the values of a random variable X would take if it were defined as number of tails when two coins are tossed simultaneously?

Solutions:

Sample Space of the experiment (Tossing of two coins simultaneously) is,
$S / \Omega = \{ TT, TH, HT, HH \}$

Let X be the number of tails obtained tossing two coins.
$X : \Omega \rightarrow R$
X(TT)=2, X(TH)=1, X(HT)=1, X(HH)=0

Since the random variable X is a number of tails, it takes three distinct values
{ 0, 1, 2 }

Remark: Several random variables can be defined on the same sample space **Ω**. For example in the Example1, one can define Y= Number of heads or Z= Difference between number of heads and number of tails.

Following are some of the examples of discrete variable.

a) Number of days of rainfalls in Mumbai.

b) Number of patients cured by using certain drug during pandemic.

c) Number of attempts required to pass the exam.

d) Number of accidents on a sea link road.

e) Number of customers arriving at shop.

f) Number of students attending class.

Definition 3: Let X be a discrete random variable defined on a sample space **Ω**. Since **Ω** contains either finite of countable infinite elements, and X is a function on **Ω**, X can take either finite or countably infinite values. Suppose X takes values x_1, x_2 $x_{3,....}$ then the set $\{ x_1, x_2 x_{3,...} \}$ is called the range set of X.

In Example1, the range set of X = Number of tails is { 0,1,2}

6.2.2 Continuous Random Variable:

A sample space which is finite or countably infinite is called as denumerable of countable. If the sample space is not countable then it is called **continuous.** For a continuous sample space Ω we can not have one to one correspondence between Ω and set of natural numbers { 1,2,…}

Random variable could also be such that their set of possible values is uncountable. Examples of such random variables are time taken between arrivals of two vehicles at petrol pump or life in hours of an electrical component.

In general if we define a random variable X(ω) as a real valued function on domain Ω. If the range set of X(ω) is continuous, the random variable is continuous. The range set will be a subset of real line.

Following are some of the examples of continuous random variable

a) Daily rainfall in mm at a particular place.

b) Time taken for an angiography operation at a hospital

c) Weight of a person in kg.

d) Height of a person in cm.

e) Instrumental error (measured in suitable units) in the measurement.

6.2.3 Distinction between continuous random variable and discrete random variable:

1) A continuous random variable takes all possible values in a range set. The ser is in the form of interval. On the other hand discrete random variable takes only specific or isolated values.

2) Since a continuous random variable takes uncountably infinite values, no probability mass can be attached to a particular value of random variable X. Therefore, P(X = x) = 0, for all x. However in case of a discrete random variable, probability mass is attached to individual values taken by random variable. In case of continuous random variable probability is attached to an interval which is a subset of R.

6.3 Probability Distributions of Discrete Random Variable

Each outcome i of an experiment have a probability P (i) associated with it. Similarly every value of random variable X= x_i is related to the outcome i of an experiment. Hence, for every value of random variable x_i, we have a unique real

value P(i) associated . Thus, every random variable X has probability P associated with it. This function P (X= x_i) from the ser of all events of the sample space Ω is called a probability distribution of the random variable.

The probability distribution (or simply distribution) of a random variable X on a sample space Ω is set of pairs (X= x_i, P (X= x_i)) for all x_i : ∈ x(Ω), where P (X= x_i) is the probability that X takes the value x_i .

Consider the experiment of tossing two unbiased coins simultaneously. X= number of tails observed in each tossing. Then range set of X = { 0, 1,2 }. Although we cannot in advance predict what value X will take, we can certainly state the probabilities with which X will take the three values 0,1,2. The following table helps to determine such probabilities.

Outcome	Probability of Outcome	Value of X
HH	1/4	0
TH	1/4	1
HT	1/4	1
TT	1/4	2

Following events are associated with the distinct values of X

(X = 0) ⇒ { HH}

(X = 1) ⇒ { TH,HT}

(X = 2) ⇒ { TT}

Therefore, probabilities of various values of X are nothing but the probabilities of the events with which the respective values are associated.

P (X = 0) ⇒ P { HH} = 1/4

P (X = 1) ⇒ P { TH, HT} = 1/4 + 1/4 = 1/2

P (X = 2) ⇒ P {TT} = 1/4

Example2:

A random variable is number of tails when a coin is flipped thrice. Find probability distribution of the random variable.

Solution:

Sample space Ω = { HHH, THH, HTH, HHT, TTH, THT, HTT, TTT }

The required probability distribution is

Value of Random Variable	**$X = x_i$**	**0**	**1**	**2**	**3**
Probability	**$P(X = x_i)$**	1/8	3/8	3/8	1/8

Example 3:

A random variable is sum of the numbers that appear when a pair of dice is rolled. Find probability distribution of the random variable.

Solution:

X(1,1) = 2, X(1, 2) = X(2, 1) = 3, X(1, 3) = X(3, 1) = X (2, 2) = 4 etc..;

The probability distribution is,

$X = x_i$	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
$P(X = x_i)$	1/36	2/36	3/36	4/36	5/36	6/35	5/36	4/36	3/36	2/36	1/36

6.3.1 Probability Mass Function (p.m.f.)

Let X be a discrete random variable defined on a sample space Ω / S . Suppose $\{x_1, x_2, x_3, \ldots\ldots x_n\}$ is the range set of X. With each of x_i, we assign a number $P(x_i) = P(X = x_i)$ called the probability mass function (p.m.f.) such that,

$P(x_i) \geq 0$ for $i = 1, 2, 3 \ldots,n$ and (AND)

$\sum_{i=1}^{n} P(x_i) = 1$

The table containing, the value of X along with the probabilities given by probability mass function (p.m.f.) is called probability distribution of the random variable X.

For example,

$X = x_i$	**X_1**	**X_2**	…..	**…..**	**X_i**	…	**….**	**X_n**	**Total**
$P(X = x_i)$	P_1	P_2	….	….	P_i	…	…	P_n	1

Remark: Properties of a random variable can be studied only in terms of its p.m.f. We need not have refer to the underlying sample space Ω, once we have the probability distribution of random variable.

Example 4:

A fair die is rolled and number on uppermost face is noted. Find its probability distribution (p.m.f.)

Solution:

X= Number on uppermost face.

Therefore, Range set of X = {1, 2, 3, 4, 5, 6 }

Probability of each of the element is = 1 / 6

The Probability distribution of X is

$X = x_i$	1	2	3	4	5	6	Total
$P(X = x_i)$	1/ 6	1/ 6	1/ 6	1/ 6	1/ 6	1/ 6	1

Example 5:

A pair of fair dice is thrown and sum of numbers on the uppermost faces is noted. Find its probability distribution (p.m.f.).

Solution:

X = Sum of numbers on the uppermost faces.

Ω contains 36 elements (ordered pairs)

Range set of X = { 2, 3, 4, 5, 6, 7, 8,9, 10, 11, 12 }

Since, X (1, 1) = 2 and X (6, 6) = 12

Value of X	Subset of Ω	$P_i = P (X = i)$
2	{ (1,1) }	1/36
3	{ (1,2), (2,1) }	2/36
4	{ (1,3), (2,2), (3,1) }	3/36
5	{ (1,4), (2,3), (3,2), (4,1) }	4/36
6	{ (1,5), (2,4), (3,3), (4,2), (5,1) }	5/36
7	{ (1,6), (2,5), (3,4), (4,3), (5,2), (6,1) }	6/36
8	{ (2,6), (3,5), (4,4), (5,3), (6,2) }	5/36
9	{ (3,6), (4,5), (5,4), (6,3) }	4/36
10	{ (4,6), (5,5), (6,4) }	3/36
11	{ (5,6) (6,5) }	2/36
12	{ (6,6) }	1/36

Example 6:

Let X represents the difference between the number of heads and the number of tails obtained when a fair coin is tossed three times. What are the possible values of X and its p.m.f.?

Solution:
Coin is fair. Therefore, probability of heads in each toss is $P_H = 1/2$. Similarly, probability of tails in each toss is $P_T = 1/2$.

X can take values n – 2r where n=3 and r = 0, 1, 2, 3.

e.g. X = 3 (HHH), X = 1 (HHT, HTH, THH), X = -1 (HTT, THT, HTT) and X = -3 (TTT)

Thus the probability distribution of X (possible values of X and p.m.f) is

$\mathbf{X = x_i}$	-3	-1	1	3	Total
p.m.f. $P(X = x_i)$	1/ 8	3/ 8	3/ 8	1/ 8	1

Example 7:

Let X represents the difference between the number of heads and the number of tails obtained when a coin is tossed n times. What are the possible values of X ?

Solution:

When a coin is tossed n times, number of heads that can be obtained are n, n-1, n-2, ……,2, 1, 0. Corresponding number of tails are 0, 1, 2,……, n-2, n-1, n. Thus the sum of number of heads and number of tails must be equal to number of trials n.

Hence, values of X are from n to –n as n, n-2, n-4, ……, n-2r

where r = 0,1, 2, 3, ….., n

Note if n is even X has one of its value as zero also. However, if n is odd X has values -1, 1 but not zero.

6.4 Probability Distributions of Continuous Random Variable

In case of discrete random variable using p.m.f. we get probability distribution of random variable, however in case of continuous random variable probability mass is not attached to any particular value. It is attached to an interval. The probability attached to an interval depends upon its location.

For example, P (a < X < b) varies for different values of a and b. In other words, it will not be uniform. In order to obtain the probability associated with any interval, we need to take into account the concept of probability density.

6.4.1 Probability Density Function (p.d.f.)

Let X be a continuous random variable. Function *f(x)* defined for all real $x \in (-\infty, \infty)$ is called probability density function p.d.f. if for any set B of real numbers, we get probability,

$$P\{X \in B\} = \int_B f(x)dx$$

All probability statements about X can be answered in terms of *f (x).* Thus,

$$P\{a \leq X \leq B\} = \int_a^b f(x)dx$$

Note that probability of a continuous random variable at any particular value is zero, since

$$P\{X = a\} = P\{a \leq X \leq a\} = \int_a^a f(x)dx = 0$$

6.5 Properties of Random variable and their probability distributions

Properties of a random variable can be studied only in terms of its p.m.f . or p.d.f. We need not have refer to the underlying sample space Ω, once we have the probability distribution of random variable. Since the random variable is a function relating all outcomes of a random experiment, the probability distribution of random variable must satisfy Axioms of probability. These in case of discrete and continuous random variable are stated as,

Axiom I : Any probability must be between zero and one.

For discrete random variable : $0 \leq p(x_i) \leq 1$

For continuous random variable : For any real number a and b

$0 \leq P\{a \leq x \leq b\} \leq 1$ **OR** $0 \leq \int_a^b f(x)dx \leq 1$

Axiom II : Total probability of sample space must be one

For discrete random variable : $\sum_{i=1}^{\infty} p(x_i) = 1$

For continuous random variable : $\int_{-\infty}^{\infty} f(x)dx = 1$

Axiom III : For any sequence of mutually exclusive events $E_1, E_2, E_3, \ldots\ldots$ i.e

$E_i \cap E_j = \Phi$ for $i \neq j$, probability of a union set of events is sum of their individual probabilities. This axiom can also be written as $P(E1 \cup E2) = P(E_1) + P(E_2)$ where E_1 and E_2 are mutually exclusive events.

For discrete random variable : $\boldsymbol{P}\left(\bigcup_{\mathbf{i=1}}^{\infty} \boldsymbol{E_i}\right) = \sum_{i=1}^{\infty} P(E_i)$

For continuous random variable : $\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$ also,

$P(a \leq x \leq b \cup c \leq x \leq d) = \int_a^b f(x)dx + \int_c^d f(x)dx$

Axiom IV : $P(\Phi) = 0$

6.6 Cumulative Distribution Function (c.d.f.)

It is also termed as just a distribution function. It is the accumulated value of probability up to a given value of the random variable. Let X be a random variable, then the cumulative distribution function (c.d.f.) is defined as a function F(a) such that,

$F(a) = P\{X \leq a\}$

6.6.1 Cumulative Distribution Function (c.d.f.) for Discrete Random Variable

Let X be a discrete random variable defined on a sample space S taking values $\{x_1, x_2, \ldots, x_n\}$

With probabilities $p(x_i)$, $p(x_2)$, …. $p(x_n)$ respectively. Then cumulative distribution function (c.d.f.) denoted as F(a) and expressed in term of p.m.f. as

$$F(a) = \sum_{x_i \leq a} p(x_i)$$

Note:

1. The c.d.f. is defined for all values of $x_i \in R$. However, since the random variable takes only isolated values, c.d.f. is constant between two successive values of X and has steps at the points $x_i I$, i = 1, 2, …, n. Thus, the c.d.f for a discrete random variable is a step function.
2. $F(\infty) = 1$ and $F(-\infty) = 0$

Properties of a random variable can be studied only in term of its c.d.f. We need not refer to the underlying sample space or p.m.f., once we have c.d.f. of a random variable.

6.5.2 Cumulative Distribution Function (c.d.f.) for Continuous Random Variable

Let X be a continuous random variable defined on a sample space S which p.d.f. *f(x)*. Then cumulative distribution (c.d.f.) denoted as F (a) and expressed in term of p.d.f. as,

$$F(a) = \int_{-\infty}^{\infty} f(x)dx$$

Also, differentiating both sides we get,

$\frac{d}{da}F(a) = f(a)$, thus the density is the derivative of the cumulative distribution function.

6.7 Properties of Cumulative Distribution Function (c.d.f.)

1. F (x) is defined for all x € R, real number.
2. $0 \leq F(x) \leq 1$
3. F (x) is a non-decreasing function of x. [if a < b,then F (a) ≤ F (b).
4. F (∞) = 1 and F (-∞) = 0

   Where $F(-\infty) = \lim_{x \to -\infty} F(x)$, $F(\infty) = \lim_{x \to \infty} F(x)$
5. Let a and b be two real numbers where a < b ; then using distribution function, we can compute probabilities of different events as follows.

   i) $P(a < X \leq b) = P[X \leq b] - P[X \leq a]$
   $= F(b) - F(a)$

   ii) $P(a \leq X \leq b) = P[X \leq b] - P[X \leq a] + P(X=a)$
   $= F(b) - F(a) + P(a)$

   iii) $P(a \leq X < b) = P[X \leq b] - P[X \leq a] - P(X=b) + P(X=a)$
   $= F(b) - F(a) - P(b) + P(a)$

   iv) $P(a < X < b) = P[X \leq b] - P[X \leq a] - P(X=b)$
   $= F(b) - F(a) - P(b)$

v) $P(X > a) = 1 - P[X \leq a] = 1 - F(a)$

vi) $P(X \geq a) = 1 - P[X \leq a] + P[X=a] = 1 - F(a) + P(a)$

vii) $P(X=a) = F(a) - \lim_{n\to\infty} F\left(a - \frac{1}{n}\right)$

viii) $P(X < a) = \lim_{n\to\infty} \left(a - \frac{1}{n}\right)$

Example 8:

The following is the cumulative distribution function of a discrete random variable.

$X = x_i$	-3	-1	0	1	2	3	5	8
F(x)	0.1	0.3	0.45	0.65	0.75	0.90	0.95	1.00

i) Find the p.m.f of X

ii) $P(0 < X < 2)$

iii) $P(1 \leq X \leq 3)$

iv) $P(-3 < X \leq 2)$

v) $P(1 \leq X < 1)$

vi) P(X = even)

vii) $P(X > 2)$

viii) $P(X \geq 3)$

Solution:

i) Since $F(x_i) = \sum_{j=1}^{i} P_j$

$F(x_{i-1}) = \sum_{j=1}^{i-1} P_j$

$\therefore P_i = \sum_{j=1}^{i} P_j - \sum_{j=1}^{i-1} P_j = F(x_i) - F(x_{i-1})$

$\therefore$ The p.m.f of X is given by

$X = x_i$	-3	-1	0	1	2	3	5	8
F(x)	0.1	0.2	0.15	0.2	0.1	0.15	0.05	0.05

ii) $P(0 < X < 2) = F(2) - F(0) - P(2) = 0.75 - 0.45 - 0.1 = 0.2$

iii) $P(1 \leq X \leq 3) = F(3) - F(1) + P(1) = 0.9 - 0.65 + 0.2 = 0.45$

iv) $P(-3 < X \leq 2) = F(2) - F(-3) = 0.75 - 0.1 = 0.65$

v) $P(1 \leq X < 1) = F(1) - F(-1) - P(1) + P(-1) = 0.65 - 0.3 - 0.2 + 0.2 = 035$

vi) $P(X = \text{even}) = P(x=0) + P(x=2) + P(x=8) = 0.15 + 0.1 + 0.05 = 0.3$

vii) $P(X > 2) = 1 - F(2) = 1 - 0.75 = 0.25$ OR
$= P(x=3) + P(x=5) + P(x=8) = 0.15 + 0.05 + 0.05 = 0.25$

viii) $P(X \geq 3) = 1 - F(3) + P(3) = 1 - 0.9 + 0.15 = 0.25$

Example 9:

A random variable has the following probability distribution

Values of X	0	1	2	3	4	5	6	7	8
P(x)	a	$3a$	$5a$	$7a$	$9a$	$11a$	$13a$	$15a$	$17a$

(1) Determine the value of a

(2) Find (i) $P(x < 3)$ (ii) $P(x \leq 3)$ (iii) $P(x > 7)$ (iv) $P(2 \leq x \leq 5)$ (v) $P(2 < x < 5)$

(3) Find the cumulative distribution function of x.

Solution:

1. Since p_i is the probability mass function of discrete random variable X,
   We have $\Sigma p_i = 1$
   $\therefore a + 3a + 5a + 7a + 9a + 11a + 13a + 15a + 17a = 1$
   $81a = 1$
   $a = 1/81$

2. (i) $P(x < 3) = P(x=0) + P(x=1) + P(x=2)$
   $= a + 3a + 5a$
   $= 9a = 9 * (1 / 81) = 1 / 9$

   (ii) $P(x \leq 3) = P(x=0) + P(x=1) + P(x=2) + P(x=3)$
   $= a + 3a + 5a + 7a$
   $= 16a = 16 * (1 / 81) = 16 / 81$

   (iii) $P(x > 7) = P(x = 8) = 17a = 17 * (1 / 81) = 17 / 81$

   (iv) $P(2 \leq x \leq 5) = P(x=2) + P(x=3) + P(x=4) + P(x=5)$
   $= 5a + 7a + 9a + 11a = 32a = 32 * (1 / 81) = 32 / 81$

   (v) $P(2 < x < 5) = P(x = 3) + P(x = 4) = 7a + 9a = 16a = 16 * (1 / 81) = 16 / 81$

3. The distribution function is as follows:

X=x	0	1	2	3	4	5	6	7	8
F(x)= P (X≤ x)	a	$4a$	$9a$	$16a$	$25a$	$36a$	$49a$	$64a$	$81a$
(or) F(x)	$\frac{1}{81}$	$\frac{4}{81}$	$\frac{9}{81}$	$\frac{16}{81}$	$\frac{25}{81}$	$\frac{36}{81}$	$\frac{49}{81}$	$\frac{64}{81}$	$\frac{81}{81}=1$

Example 10:

Find the probability between X = 1 and 2 i.e. P(1 ≤ X ≤ 2) for a continuous random variable whose p.d.f. is given as

$f(x) = (\frac{1}{6}x + k)$ for $0 \le x \le 3$

Solution:

Now, p.d.f must satisfy the probability Axiom II. Thus

$\int_{-\infty}^{\infty} f(x)dx = 1$

$\int_0^3 \left(\frac{1}{6}x + k\right)dx = \left[\frac{x^2}{12} + kx\right]_0^3 = \left[\frac{3^2}{12} + 3k\right] + 0 = 1$

$\therefore 12k = 1$

$\therefore k = \frac{1}{12}$

Now, P (1 ≤ X ≤ 2) = $\int_1^2 f(x)dx = \int_1^2 \left(\frac{1}{6}x + \frac{1}{12}\right)dx = \frac{1}{3}$

Example 11:

A continuous random variable whose p.d.f. is given as

$$f(x) = \begin{cases} kx(2-x) & 0 < x < 2 \\ \mathbf{0} & otherwise \end{cases}$$

i) Find k

ii) Find P ($x < \frac{1}{2}$)

Solution:

i) Now, p.d.f must satisfy the probability Axiom II. Thus

$\int_{-\infty}^{\infty} f(x)dx = 1$

$\int_0^2 kx(2-x)dx = \left[kx^2 - \frac{kx^3}{3}\right]_0^2 = 1$

$\therefore k = \frac{3}{4}$

ii)

Now, P $(x < \frac{1}{2}) = \int_{-\infty}^{1/2} f(x)dx$

$$= \int_{-\infty}^{\frac{1}{2}} \frac{3}{4}x(2-x)dx \quad = \left[\frac{3}{4}x^2 - \frac{x^3}{3}\right]_0^{1/2} = \frac{5}{12}$$

Example 12:

A random variable is a number of tails when a coin is tossed three times. Find p.m.f. and c.d.f of the random variable.

Solution:

S / Ω = { TTT, HTT, THT, TTH, HHT, HTH,THH, HHH }

n(Ω) = 8

X = x_i	0	1	2	3	Total
p.m.f. P(X = x_i)	1/ 8	3/ 8	3/ 8	1/ 8	1
c.d.f F(a) = P{ X = $x_i \leq a$}	1/ 8	4 / 8	7 / 8	1	

c.d.f. will be describe as follows:

$$\begin{aligned} F(a) &= 0 && -\infty < a < 0 \\ &= \frac{1}{8} && 0 \leq a < 1 \\ &= \frac{4}{8} && 1 \leq a < 2 \\ &= \frac{7}{8} && 2 \leq a < 3 \\ &= 1 && 3 \leq a < -\infty \end{aligned}$$

Example 13:

A c.d.f. a random variable is as follows

$$\begin{aligned} F(a) &= 0 && -\infty < a < 0 \\ &= \frac{1}{2} && 0 \leq a < 1 \\ &= \frac{2}{3} && 1 \leq a < 2 \\ &= \frac{11}{12} && 2 \leq a < 3 \\ &= 1 && 3 \leq a < -\infty \end{aligned}$$

Find i) P($X < 3$) ii) P ($X = 1$)

Solution:

i) $P(X < 3) = \lim_{n\to\infty} F\left(3 - \frac{1}{n}\right) = \frac{11}{12}$

ii) $P(X = 1) = P(X \leq 1) - P(X < 1)$

$= F(1) - \lim_{n\to\infty} F\left(3 - \frac{1}{n}\right) = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$

6.8 Expectation or Expected Value of Random Variable :

Expectation is a very basic concept and is employed widely in decision theory, management science, system analysis, theory of games and many other fields. It is one of the most important concepts in probability theory of a random variable.

Expectation of X is denoted by E(X). The expected value or mathematical expectation of a random variable X is the weighted average of the values that X can assume with probabilities of its various values as weights. Expected value of random variable provides a central point for the distribution of values of random variable. So expected value is a mean or average value of the probability distribution of the random variable and denoted as 'μ' (read as ' mew'). Mathematical expectation of a random variable is also known as its arithmetic mean.

$\mu = E(X)$

6.8.1 Expected Value of a Discrete Random Variable :

If X is a discrete random variable with p.m.f. $P(x_i)$, the expectation of X, denoted by E(X), is defined as,

$$E(X) = \sum_{i=1}^{n} x_i * P(x_i) \qquad \text{Where } x_i \text{ for i = 1, 2, ...n (values of } X)$$

6.8.2 Expected Value of a Continuous Random Variable :

If X is a discrete random variable with p.d.f. f(x), the expectation of X, denoted by E(X), is defined as,

$$E(X) = \int_{-\infty}^{\infty} f(x)dx$$

6.8.3 Properties of Expectation

1. For two random variable X and Y if E(X) and E(Y) exist, E(X + Y) = E(X) + E(Y) . This is known as addition theorem on expectation.

expectation.

3. The expectation of a constant is the constant it self. ie E(C) = C

4. E(cX) = cE(X)

5. E(aX+b) = aE(X) +b

6.9 Expectation of a Function

Let Y = g (X) is a function of random variable *X*, then *Y* is also a random variable with the same probability distribution of *X*.

For discrete random variable *X* and *Y,* probability distribution of *Y* is also P(x_i). Thus the expectation of *Y* is,

$$E(Y) = E[(g(x_i)] = \sum_{i=1}^{n} g(x_i) * P(x_i)$$

For continuous random variable *X* and *Y,* probability distribution of *Y* is also f(x). Thus the expectation of *Y* is,

$$E(Y) = E[(g(x_i)] = \int_{-\infty}^{\infty} g(x) * f(x)dx$$

Example 14:

A random variable is number of tails when a coin is tossed three times. Find expectation (mean) of the random variable.

Solution:

S / Ω = { TTT, HTT, THT, TTH, HHT, HTH,THH, HHH }

n(Ω) = 8

$\mathbf{X = x_i}$	0	1	2	3
p.m.f. P(X = x_i)	1/ 8	3/ 8	3/ 8	1/ 8
$\boldsymbol{x_i * P(x_i)}$	0	3 / 8	6 / 8	3 / 8

$E(X) = \sum_{i=1}^{4} x_i * P(x_i)$ Where x_i for i = 1, 2, …..n (values of *X)*

$E(X) = 0+\frac{3}{8}+\frac{6}{8}+\frac{3}{8}=\frac{12}{8}=\frac{3}{2}$

Example 15:

X is random variable with probability distribution

$X = x_i$	0	1	2
p.m.f. $P(X = x_i)$	0.3	0.3	0.4

Y = g(X) = 2X + 3

Find expected value or mean of Y . (i.e. E(*Y*))

Solution:

Y = g(X) = 2X + 3

When X=0, Y= ???

Y = 2X + 3 = 2(0) + 3 = 3,

Similarly,

when X = 1, Y = 5, when X = 2, Y = 7

$X = x_i$	0	1	2
$Y = y_i$	3	5	7
p.m.f. $P(Y = y_i)$	0.3	0.3	0.4

$E(Y) = E[(g(x_i)] = \sum_{i=1}^{n} g(x_i) * P(x_i) = \sum_{i=1}^{n} y_i * P(x_i)$

$E(Y) = 3\times0.3 + 5\times0.3 + 7\times0.4 = 5.2$

6.10 Variance of a Random Variable

The expected value of X (i.e. E(*X*)) provides a measure of central tendency of the probability distribution. However it does not provide any idea regarding the spread of the distribution. For this purpose, variance of a random variable is defined .

Let X be a discrete random variable on a sample space S. The variance of X denoted by *Var(X) or* σ^2 (read as 'sigma square) is defined as,

$Var(X) = E[(X - \mu)^2] = E[(X - E(X))^2] = \sum_{i=1}^{n} (x_i - \mu)^2 P(x_i)$

$Var(X) = \sum_{i=1}^{n} (x_i)^2 P(x_i) - 2\mu \sum_{i=1}^{n} x_i P(x_i) + \mu^2 \sum_{i=1}^{n} P(x_i)$

$= E(X^2) - 2\mu E(X) + \mu^2$ ……..

$[E(X) = \sum_{i=1}^{n} x_i P(x_i) \ \ and \sum_{i=1}^{n} P(x_i) = 1]$

$= E(X^2) - 2\mu.\ \mu + \mu^2$

$= E(X^2) - \mu^2 = E(X^2) - [E(X)]^2$

For continuous random variable,

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x)dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)dx - \int_{-\infty}^{\infty} 2\mu x f(x)dx + \int_{-\infty}^{\infty} \mu^2 f(x)dx$$

$= E(X^2) - 2\mu\ E(X) - [E(X)]^2$

$= E(X^2) - 2\mu.\ \mu + \mu^2$

$= E(X^2) - \mu^2 = E(X^2) - [E(X)]^2$

Since dimensions of variance are square of dimensions of X, foe comparison, it is better to take square root of variance. It is known as standard deviation and denoted be S.D.(X) or σ (sigma)

S.D.= $\sigma = \sqrt{Var(x)}$

6.10.1 Properties of Variance:

1. Variance of constant is zero. ie Var(c) = 0
2. Var(X+c) = Var X

   Note: This theorem gives that variance is independent of change of origin.
3. Var (aX) = a^2 var(X)

   Note: This theorem gives that change of scale affects the variance.
4. Var (aX+b) = a^2Var(X)
5. Var (b-ax) = a^2 Var(x)

Example 16:

Calculate the variance of X, if X denote the number obtained on the face of fair die.

Solution:

X is random variable with probability distribution

$\mathbf{X = x_i}$	1	2	3	4	5	6
p.m.f. P(X = x_i)	1/6	1/6	1/6	1/6	1/6	1/6
$\boldsymbol{x_i * P(x_i)}$	1/6	2/6	3/6	4/6	5/6	6/6
$\boldsymbol{x^2{}_i * P(x_i)}$	1/6	4/6	9/6	16/6	25/6	36/6

$E(X) = \sum_{i=1}^{6} x_i * P(x_i)$ Where x_i for i = 1, 2, 3, 4, 5,6 (values of *X)*

$E(X) = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = \frac{21}{6} = 3.5$

$E(X^2) = \sum_{i=1}^{6} x^2{}_i * P(x_i)$ Where x_i for i = 1, 2, 3, 4, 5,6 (values of *X)*

$E(X^2) = \frac{1}{6} + \frac{4}{6} + \frac{9}{6} + \frac{16}{6} + \frac{25}{6} + \frac{36}{6} = \frac{91}{6}$

$\sigma^2 = \text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{105}{36} = 2.9167$

Example 17:

Obtain variance of r.v. X for the following p.m.f.

$\mathbf{X = x_i}$	0	1	2	3	4	5
p.m.f. P(X = x_i)	0.05	0.15	0.2	0.5	0.09	0.01

Solution:

X is random variable with probability distribution

$\mathbf{X = x_i}$	0	1	2	3	4	5
p.m.f. P(X = x_i)	0.05	0.15	0.2	0.5	0.09	0.01
$\boldsymbol{x_i * P(x_i)}$	0	0.15	0.40	1.50	0.36	0.05
$\boldsymbol{x^2{}_i * P(x_i)}$	0	0.15	0.80	4.50	1.44	0.25

$E(X) = \sum_{i=0}^{5} x_i * P(x_i)$ Where x_i for i = 0,1, 2, 3, 4, 5 (values of *X)*

$E(X) = 2.46$

$E(X^2) = \sum_{i=0}^{5} x^2{}_i * P(x_i)$ Where x_i for i = 0,1, 2, 3, 4, 5 (values of *X)*

$E(X^2) = 7.14$

$\sigma^2 = \text{Var}(X) = E(X^2) - [E(X)]^2 = 7.14 - (2.46)^2 = 1.0884$

Example 18:

Obtain variance of r.v. X for the following probability distribution

$P(x) = \frac{x^2}{30}$, $x =$ 0,1, 2,3, 4

Solution:

X is random variable with probability distribution $P(x) = \frac{x^2}{30}$, $x =$ 0,1, 2,3, 4

$E(X) = \sum_{i=0}^{4} x_i * P(x_i)$ Where x_i for i = 0,1, 2, 3, 4 (values of *X)*

$= \sum_{i=0}^{4} x_i * \frac{x^2}{30} = \frac{1}{30} \sum_{i=0}^{4} x^3 = \frac{1}{30}(0 + 1 + 8 + 27 + 64) = \frac{100}{30} = \frac{10}{3} = 3.33$

$E(X) = \frac{10}{3}$

$E(X^2) = \frac{1}{30} \sum_{i=0}^{4} x^4$ Where x_i for i = 0,1, 2, 3, 4, 5 (values of *X)*

$E(X2) = \frac{1}{30} \sum_{i=0}^{4} x^4 = \frac{1}{30}(1 + 16 + 81 + 256) = \frac{354}{30} = 11.8$

$E(X^2) = \sigma^2 = \text{Var}(X) = E(X^2) - [E(X)]^2 = 11.8 - (3.33)^2 = 0.6889$

6.11 Summary

In this chapter, random variables, its types with its Probability Distributions, expected value and variance is discussed.

Random variable : Let S / Ω be the sample space corresponding to the outcomes of a random experiment . A function X: S → R (where R is a set of real numbers) is called as a random variable.

A random variable X is said to be discrete if it takes finite or countably infinite number of possible values. A sample space which is finite or countably infinite is called as denumerable of countable. If the sample space is not countable then it is called **continuous.**

The probability distribution (or simply distribution) of a random variable X on a sample space Ω is set of pairs (X= x_i, P (X= x_i)) for all $x_i : \in$ x(Ω), where P (X= x_i) is the probability that X takes the value x_i .

Let X be a discrete random variable defined on a sample space Ω / S . Suppose {x_1, $x_2, x_3, \ldots\ldots x_n$ } is the range set of X. With each of x_i, we assign a number $P(x_i) = P(X = x_i)$ called the probability mass function (p.m.f.) such that,

$P(x_i) \geq 0$ for $i = 1, 2, 3 \ldots, n$ and (AND)

$\sum_{i=1}^{n} P(x_i) = 1$

Probability Density Function (p.d.f.) Let X be a continuous random variable. Function *f(x)* defined for all real $x \in (-\infty, \infty)$ is called probability density function p.d.f. if for any set B of real numbers, we get probability,

$$P\{X \in B\} = \int_B f(x)dx$$

Axiom I : Any probability must be between zero and one.

Axiom II : Total probability of sample space must be one

Axiom III : For any sequence of mutually exclusive events $E_1, E_2, E_3, \ldots\ldots$ i.e

$E_i \cap E_j = \Phi$ for $i \neq j$

Cumulative Distribution Function (c.d.f.)

It is also termed as just a distribution function. It is the accumulated value of probability up to a given value of the random variable. Let X be a random variable, then the cumulative distribution function (c.d.f.) is defined as a function F(a) such that,

$$F(a) = P\{X \leq a\}$$

Expected Value of Random Variable expected value is a mean or average value of the probability distribution of the random variable $\mu = E(X)$

The variance of X denoted by *Var(X)* S.D.= $\sigma = \sqrt{Var(x)}$

6.12 Reference

Fundamentals of Mathematical Statistics S. C. Gupta, V. K. Kapoor

❖❖❖

7

MOMENTS AND MOMENT GENERATING FUNCTIONS

Unit Structure

7.1 Objectives

After going through this chapter students will be able to:

- Define moments.
- Calculate Raw movement and Central movement.
- Define Moment generating functions and calculate for random variable.
- Relation between Raw moments and Central moments.

7.2 Introduction

Any random variable X describing a real phenomenon has necessarily a bounded range of variability implying that the values of the moments determine the probability distribution uniquely. The range of variability of a random variable X is defined as the set of all observable values of X, which can occur with positive probability. Consequently, the range of any random variable of a real world quantity is finite, which further restricts the range of all the moments of X. In fact, the range of the n^{th} moment $E[X^n]$, n = 1, 2,..., reduces the range of the successive

moments $E[X^{n+k}]$, k = 1, 2,.... Thus, any available knowledge about the possible values of $E[X^n]$, can be used for drawing inference on $E[X^{n+k}]$.

The value of the first moment limits considerably the range of the second moment; etc. Thus, any knowledge about the values of lower moments may be used without a further sample for drawing inference on the higher moments.

7.3 Moments

Moments about the origin (raw moments) : The r[th] moment about the origin of a random variable X, denoted by μ'_r, is the expected value of X^r; symbolically,

$$\mu'_r = E[X^r] = \sum_x x^r \, f(x) \quad (1)$$

for r = 0, 1, 2, . . . when X is discrete and

$$\mu'_r = E[X^r] = \int_{-\infty}^{\infty} x^r \, f(x)dx \quad (2)$$

when X is continuous.

The r[th] moment about the origin is only defined if $E[X^r]$ exists. A moment about the origin is sometimes called a raw moment. Note that $\mu'_r = E[X] = \mu X$, the mean of the distribution of X, or simply the mean of X. The r[th] moment is sometimes written as function of θ where θ is a vector of parameters that characterize the distribution of X. If there is a sequence of random variables, $X_1, X_2, \ldots \ldots X_n$, we will call the r[th] population moment of the i[th] random variable $\mu'_{i,r}$ and define it as $\mu'_{i,r} = E[X_i^r]$. (3)

Central moments : The r[th] moment about the mean of a random variable X, denoted by μ_r , is the expected value of $(X - \mu X)^r$ symbolically,

$$\mu_r = E[(X - \mu X)^r] = \sum_x (x - \mu X)^r \; f(x) \quad (4)$$

for r = 0, 1, 2, . . . when X is discrete and

$$\mu_r = E[(X - \mu X)^r] = \int_{-\infty}^{\infty} (x - \mu X)^r \; f(x)dx \quad (5)$$

when X is continuous.

The r[th] moment about the mean is only defined if $E[(X - \mu X)^r]$exists. The r[th] moment about the mean of a random variable X is sometimes called the r[th] central moment of X. The r[th] central moment of X about a is defined as $E[(X - a)^r]$. If $a = \mu X$, we have the rth central moment of X about μX.

$$\mu_1 = E[(X - \mu X)^1] = \int_{-\infty}^{\infty}(x - \mu X)^1 \ f(x)dx = 0$$

$$\mu_2 = E[(X - \mu X)^2] = \int_{-\infty}^{\infty}(x - \mu X)^2 \ f(x)dx = Var(X) = \sigma^2 \qquad (6)$$

Also note that all odd moments of X around its mean are zero for symmetrical distributions provided such moments exist.

If there is a sequence of random variables, $X_1, X_2, \ldots\ldots X_n$ we will call the r[th] central population moment of the i[th] random variable $\mu_{i,r}$ and define it as

$$\mu_{i,r} = E\left(X_i^r - \mu'_{i,1}\right)^r \qquad (7)$$

When the variables are identically distributed, we will drop the i subscript and write μ'_r and μ_r.

7.4 Moment generating functions (m.g.f)

From the discussion of moments, it is apparent that moments play an important role in the characterization of various distributions. Hence to know the distribution, we need to find out the moments. For this the moment generating function is a good device. The moment generating function is a special from of mathematical expectation, and is very useful in deriving the moments of a probability distribution.

Definition: If X is a random variable, of the discrete probability distribution about the value $x = a$ then the expected value of $e^{t(x-a)}$ is known as the moment generating function, provided the expected value exists for every value of t in an interval, $-h < t < h$ where h is some positive real value. The moment generating function which is denoted as $m_a(t)$ for a discrete random variable is

$$M_a(t) = E\left(e^{t(x-a)}\right) = \sum_x e^{t(x-a)}p_x(x)$$

$$= \sum_x \left(1 + t(x-a) + \frac{t^2(x-a)^2}{2!} + \frac{t^2(x-a)^2}{3!} + \cdots.\right)p_x(x)$$

$$= 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots\ldots. = \sum_{r=0}^{\infty}\frac{t^r}{r!}\mu'_r$$

The convergence of the above sum is assumed here. In the above expression, the r[th] raw moment is the coefficient of $\frac{t^r}{r!}$ in the above expanded sum.

We find $\mu'_r = coefficient\ of\ \frac{t^r}{r!}$ in the expression of $M_a(t)$.

Otherwise differentiating $M_a(t) = 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots$ ... r-times with respect to t and then putting $t = 0$, we get

$$\mu'_r = \left[\frac{d^r}{dt^r} M_a(t)\right]_{t=0}$$

Thus the moment about any point $x = a$ can be found more conveniently using above formula.

$$M_a(t) = e^{-at}\, M_0(t)$$

Thus the moment of generating function about the point $a = e^{-at}$ (m.g.f. about the origin)

If $f(x)$ is the density function of a continuous variate X, then the moment generating function of this continuous probability distribution about $x = a$ is given by

$$M_a(t) = \int_{-\infty}^{\infty} e^{t(x-a)} f(x) dx.$$

The main use of the moment generating function is that if it exists, it can uniquely determine the distribution. But many distribution functions are known for which the moment generating functions do not exist.

The moment generating function is also useful for finding out the moments of a distribution. The technique for it is to differentiate the moment generating function with respect to t once, twice, thrice, ….and put t = 0 in the first, second, and third … derivatives to obtain the first, second, third, …. moments. From the resulting expressions we get 1st, 2nd, 3rd, …., raw momentsabout the origin. The central moments are obtained by using the relationship between raw moments and central moments.

7.5 Relation between Raw moments and Central moments

Sometimes it is convenient to convert moments about origin to central moments. The general equation for converting the nth movement about origin to the central moment is

$$\mu_n = E[(X - E(x))^n] = \sum_{j=0}^{n} \binom{n}{j} (-1)^{n-j} \mu'_j \mu^{n-j}$$

Where μ the mean of the distribution and moment about the origin is is given by

$\mu'_m = \int_{-\infty}^{\infty} x^m f(x)dx = E(x^m)$.

Now the relation between raw moments and central moment is

$\mu_2 = \mu'_2 - \mu^2$ this is commonly referred to as variance of probability distribution as we used in previous chapter $V(X) = E(X^2) - [E(X)]^2$.

For the Third central moment

$$\mu_3 = \mu'_3 - 3\mu\mu'_2 + 2\mu^3$$

For the Fourth central moment,

$\mu_4 = \mu'_4 - 4\mu\mu'_3 + 6\mu^2\mu'_2 - 3\mu^4$ and so on…

Note:

- The "zeroth" central moment μ_0 is 1.
- The first central moment μ_1 is 0. (not to be confused with the first raw moment or the expected value μ)
- The second central moment μ_2 is called the variance, and is usually denoted σ^2, where σ represents the standard deviation.
- The third and fourth central moments are used to define the standardized moments which are used to define Skewness and Kurtosis respectively.

Example 1: Find i) First fours moments about origin, ii) First fours moments about the mean, for a random variable X having probability density function

$$f(x) = 4x(9 - x^2)/81 \qquad 0 \le x \le 3$$

$$= 0 \qquad otherwise$$

Solution: For a random variable X having probability density function

$$f(x) = 4x(9 - x^2)/81 \qquad 0 \le x \le 3$$

$$= 0 \qquad otherwise$$

i) For First Fours Moments about Origin: $\mu'_r = E[X^r] = \int_{-\infty}^{\infty} x^r\, f(x)dx$

$$\mu'_1 = E(X) = \int_0^3 x\ \ 4x(9-x^2)/81dx$$

$$= \frac{4}{81}\int_0^3 x^2\,(9-x^2)dx$$

$$= \frac{4}{81}\left[\frac{9x^3}{3} - \frac{x^5}{5}\right]_0^3$$

$$= \frac{4}{81}\left[81 - \frac{243}{5}\right] = \frac{8}{5} = \mu$$

$$\mu'_2 = E(X^2) = \int_0^3 x^2\ \ 4x(9-x^2)/81dx$$

$$= \frac{4}{81}\int_0^3 x^3\,(9-x^2)dx = 3$$

$$\mu'_3 = E(X^3) = \int_0^3 x^3\ \ 4x(9-x^2)/81dx$$

$$= \frac{4}{81}\int_0^3 x^4\,(9-x^2)dx = \frac{216}{35}$$

$$\mu'_4 = E(X^4) = \int_0^3 x^4\ \ 4x(9-x^2)/81dx$$

$$= \frac{4}{81}\int_0^3 x^5\,(9-x^2)dx = \frac{27}{2}$$

ii) First fours moments about the mean(central):
Using the relation between raw moment and central moment,

$$\mu_1 = 0$$

$$\mu_2 = \mu'_2 - \mu^2 = 3 - \left(\frac{8}{5}\right)^2 = \frac{11}{25}$$

For the Third central moment

$$\mu_3 = \mu'_3 - 3\mu\mu'_2 + 2\mu^3$$

$$= \frac{216}{35} - 3(3)\left(\frac{8}{5}\right)^2 + 2\left(\frac{8}{5}\right)^3 = \frac{-32}{875}$$

For the Fourth central moment,

$$\mu_4 = \mu'_4 - 4\mu\mu'_3 + 6\mu^2\mu'_2 - 3\mu^4$$
$$= \frac{27}{2} - 4\left(\frac{216}{35}\right)\left(\frac{8}{5}\right) + 6(3)\left(\frac{8}{5}\right)^2 - 3\left(\frac{8}{5}\right)^4 = \frac{3693}{8750}$$

Example 2: Find the moment generating function of the exponential distribution

$$f(x) = \frac{1}{c}\, e^{-x/c}, \;\; 0 \le x \le \infty, \qquad c > 0.$$

Hence find its mean and standard deviation.

Solution: The moment generating function about the origin is

$$M_a(t) = \int_0^\infty e^{tx} f(x)dx.$$

$$M_0(t) = \int_0^\infty e^{tx}\frac{1}{c}\, e^{-x/c}dx.$$

$$M_0(t) = \tfrac{1}{c}\textstyle\int_0^\infty e^{(t-1/c)x}\, dx \qquad \because \; |t| < \tfrac{1}{c}$$

$$M_0(t) = \frac{1}{c}\,\frac{\left|e^{(t-1/c)x}\right|_0^\infty}{(t-1/c)}$$

$$M_0(t) = (1-ct)^{-1} = 1 + ct + c^2t^2 + \cdots \ldots \ldots$$

$$\mu'_1 = \left[\frac{d}{dt}M_0(t)\right]_{t=0} = (c + 2c^2t + 3c^3t^2 + \cdots..)_{t=0} = c$$

$$\mu'_2 = \left[\frac{d^2}{dt^2}M_0(t)\right]_{t=0} = 2c^2$$

$$\mu_2 = \mu'_2 - \left(\mu'_1\right)^2 = 2c^2 - c^2 = c^2.$$

Hence the mean is c and the standard deviation is also c.

Example 3: The random variable X can assume the values 1 and -1 with probability $\frac{1}{2}$ each. Find the moment generating function and also find the first four moments about the origin.

Solution: Let $E(e^{tx}) = e^{t(1)}\left(\frac{1}{2}\right) + e^{t(-1)}\left(\frac{1}{2}\right) = \frac{1}{2}(e^t + e^{-t})$.

By Taylor series we get,

$$e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots ..$$

$$e^{-t} = 1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + \cdots ..$$

$$\therefore \frac{1}{2}(e^t + e^{-t}) = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \cdots ..$$

Comparing with $M_a(t) = 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots$ we get

$$\mu = 0, \qquad \mu'_2 = 1, \ \mu'_3 = 0, \mu'_4 = 1$$

Therefore, the odd moments are all zero and the even moments are all one.

Example 4: A random variable X has probability density function is given by

$$f(x) = 2e^{-2x} \qquad x \geq 0$$

$$= 0 \qquad x < 0$$

Find i) the moment generating function. ii) The first four moment about the origin.

Solution: A random variable X has probability density function is given by

$$f(x) = 2e^{-2x} \qquad x \geq 0$$

$$= 0 \qquad x < 0$$

i) The moment generating function is given by

$$M_0(t) = E(e^{xt}) = \int_{-\infty}^{\infty} e^{xt} f(x)dx$$

$$= \int_{-\infty}^{\infty} e^{xt} (2e^{-2x})dx$$

$$= 2\int_{-\infty}^{\infty} e^{(t-2)x} dx$$

$$= \left[\frac{2e^{(x-2)t}}{t-2}\right]_0^{\infty}$$

$$= \frac{2}{2-t}, \quad Assuming\ t < 2$$

ii) The first four moment about the origin:

If $|t| < 2$ we have

$$\frac{2}{2-t} = \frac{1}{1-{}^{t}/_{2}} = 1 + \frac{t}{2} + \frac{t^2}{4} + \frac{t^3}{8} + \frac{t^4}{16} + \cdots \ldots$$

Comparing with $M_a(t) = 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots$ we get

$\mu = \frac{1}{2},\ \mu'_2 = \frac{1}{2},\ \ \mu'_3 = \frac{3}{4}, \mu'_4 = \frac{3}{2}$.

Example 5: The random variable X has moment generating function

$$M_x(t) = (1-2t)^{-{}^{1}/_{2}} \qquad for\ t < \tfrac{1}{2}$$

Find first and second raw moments and central moments of random variable X.

Solution: The given moment generating function of random variable X is

$$M_x(t) = (1-2t)^{-{}^{1}/_{2}} \qquad for\ t < \frac{1}{2}$$

differentiating $M_x(t) = 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots \ldots$ *r*-times with respect to *t* and then putting *t = 0* , we get

$$\mu'_r = \left[\frac{d^r}{dt^r} M_x(t)\right]_{t=0}$$

For r = 1 we get,

$$\mu'_1 = \left[\frac{d}{dt} M_x(t)\right]_{t=0}$$

$$= \left[\frac{d}{dt}(1-2t)^{-{}^{1}/_{2}}\ \right]_{t=0}$$

$$= \frac{-1}{2}(1-2t)^{-{}^{3}/_{2}}(-2)$$

$$= \ (1-2t)^{-{}^{3}/_{2}} = 0$$

For r = 2 we get

$$\mu'_2 = \left[\frac{d^2}{dt^2} M_x(t)\right]_{t=0}$$

$$= \left[\frac{d^2}{dt^2}(1-2t)^{-1/2}\right]_{t=0}$$

$$= \frac{-3}{2}(1-2t)^{-5/2}(-2)$$

$$= \ 3(1-2t)^{-5/2} = 3$$

Example 6: Suppose that you have fair die, and let X be the random variable representing the value of the number rolled.

i) Write down the moment generating function for X.
ii) Use the moment generating function to compute first and second moment of X.

Solution: The probability of getting each number is $\frac{1}{6}$.

$$M_x(t) = E[e^{tx}]$$

$$M_x(t) = \frac{1}{6}e^{1\cdot t} + \frac{1}{6}e^{2\cdot t} + \frac{1}{6}e^{3\cdot t} + \frac{1}{6}e^{4\cdot t} + \frac{1}{6}e^{5\cdot t} + \frac{1}{6}e^{6\cdot t}$$

$$M_x(t) = \frac{1}{6}[e^{t} + e^{2t} + e^{3t} + e^{4t} + e^{5t} + e^{6t}]$$

differentiating $M_x(t) = 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots$ ... *r*-times with respect to *t* and then putting *t = 0* , we get

$$\mu'_r = \left[\frac{d^r}{dt^r} M_x(t)\right]_{t=0}$$

For r = 1 we get,

$$\mu'_1 = \left[\frac{d}{dt} M_x(t)\right]_{t=0}$$

$$\mu'_1 = \left[\frac{d}{dt}\frac{1}{6}[e^{t} + e^{2t} + e^{3t} + e^{4t} + e^{5t} + e^{6t}]\right]_{t=0}$$

$$\mu'_1 = \frac{1}{6}\left[[e^{t} + 2e^{2t} + 3e^{3t} + 4e^{4t} + 5e^{5t} + 6e^{6t}]\right] = \frac{7}{2}$$

For r = 2 we get

$$\mu'_2 = \left[\frac{d^2}{dt^2} M_x(t)\right]_{t=0}$$

$$= \left[\frac{d^2}{dt^2}\frac{1}{6}[e^t + e^{2t} + e^{3t} + e^{4t} + e^{5t} + e^{6t}]\right]_{t=0}$$

$$\mu'_2 = \frac{1}{6}\left[[e^t + 4e^{2t} + 9e^{3t} + 16e^{4t} + 25e^{5t} + 36e^{6t}]\right]_{t=0} = \frac{91}{6}$$

7.6 Lets sum up

In this chapter we have learnt the following:

- Definition of raw moment or moment about origin.
- Definition of central moment or moment about mean.
- Calculating raw and central moment.
- Relation between raw moment and central moment.
- Moment generating function for random variable X.

7.7 Unit End exercise

1. Find moment of generating function of the random variable X take values -1/2 and 1/2 with probability ½ each. Also find the first four moment about origin.

2. Find moment of generating function of the random variable X having probability density function

$$f(x) = \frac{x}{2} \qquad 0 \le x \le 2$$

$$= 0 \qquad otherwise$$

   Also find the first four moment about origin.

3. Find the first four moments about mean for random variable X having probability density function

$$f(x) = 5x^2 \qquad 0 \le x \le 1$$

$$= 0 \qquad otherwise$$

4. The random variable X has moment generating function

$$M_x(t) = (1-3t)^{1/2} \qquad for\ t < \frac{1}{3}$$

Find first and second raw moments and central moments of random variable X.

5. Let X be random variable whose probability density function is given by

$$f(x) = e^{-2x} + \frac{1}{2}e^{-x} \qquad x > 0$$

$$= 0 \qquad otherwise$$

   i) Write down the moment generating function for X.
   ii) Use the moment generating function to compute first and second moment of X.

6. Consider a random variable with two possible values 0 and 1, and corresponding probability p and (1- p) respectively. Find moment generating function.

7. Find the moment generating function for discrete random variable X whose probability mass function is

$$P_X(k) = \frac{1}{3} \qquad for\ k = 1$$

$$= \frac{2}{3} \qquad for\ \ k = 2$$

8. Find the moment generating function of the exponential distribution

$$f(x) = \frac{1}{3}\, e^{-x},\ \ 0 \le x \le \infty,$$

   Hence find its mean and standard deviation.

9. Suppose that you have three coins are tossed, and let X be the random variable representing the value of the number heads.

   i) Write down the moment generating function for X.

   ii) Use the moment generating function to compute first and second moment of X.

10. Suppose that you have 4 face die, and let X be the random variable representing the value of the number rolled.

    i) Write down the moment generating function for X.
    ii) Use the moment generating function to compute first and second moment of X
    iii) Also find mean , variance and standard deviation.

7.8 Reference

Fundamentals of Mathematical Statistics by S. C. Gupta, V. K. Kapoor

Basic statistics by B. L. Agrawal

❖❖❖

8

DISTRIBUTIONS: DISCRETE DISTRIBUTIONS

Unit Structure

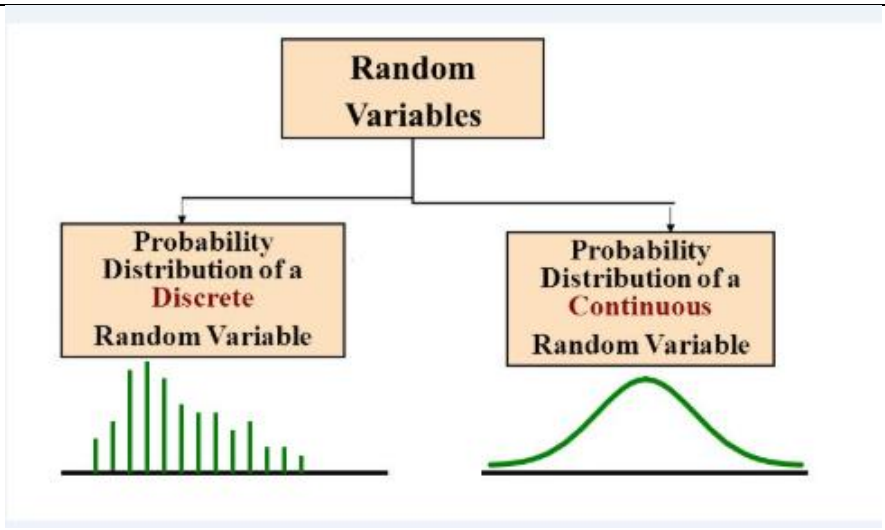
8.0 Objectives

After going through this unit, you will be able to:

- Understand the need of standard probability distribution as models
- Learn the Probability distributions and compute probabilities
- Understand the specific situations for the use of these models
- Learn interrelations among the different probability distributions.

8.1 Introduction

In previous unit we have seen the general theory of univariate probability distributions. For a discrete random variable, p.m.f. can be calculated using underlying probability structure on the sample space of the random experiment. The p.m.f. can be expressed in a mathematical form. This probability distribution can be applied to variety of real-life situations which possess some common features. Hence these are also called as 'Probability Models'.



Discrete distributions:	**Continuous distributions:**
Uniform, Bernoulli, Binomial, Poisson	Uniform distributions, Exponential

In this unit we will study some probability distributions. Viz. Uniform, Binomial, Poisson and Bernoulli distributions.

8.2 Uniform Distribution

Uniform distribution is the simplest statistical distribution. When a coin is tossed the likelihood of getting a tail or head is the same. A good example of a discrete uniform distribution would be the possible outcomes of rolling a 6-sided fair die. Ω = { 1, 2, 3, 4, 5, 6 }

In this case, each of the six numbers has an equal chance of appearing. Therefore, each time the fair die is thrown, each side has a chance of 1/6. The number of values is finite. It is impossible to get a value of 1.3, 4.2, or 5.7 when rolling a fair die. However, if another die is added and they are both thrown, the distribution that results is no longer uniform because the probability of the sums is not equal.

A deck of cards also has a uniform distribution. This is because an individual has an equal chance of drawing a spade, a heart, a club, or a diamond i.e. 1/52.

Consider a small scale company with 30 employees with employee id 101 to 130. A leader for the company to be selected at random. Therefore a employee id is selected randomly from 101 to 130. If X denotes the employee id selected then since all the id's are equally likely, the p.m.f. of X is given by,

$$P(x) = \frac{1}{30}; \qquad x = 101, 102, \ldots\ldots 150$$

$$= 0 \qquad \text{otherwise}$$

Such distribution is called as a discrete uniform distribution. The discrete uniform distribution is also known as the "equally likely outcomes" distribution.

The number of values is finite. It is impossible to get a value of 1.3, 4.2, or 5.7 when rolling a fair die. However, if another die is added and they are both thrown, the distribution that results is no longer uniform because the probability of the sums is not equal. Another simple example is the probability distribution of a coin being flipped. The possible outcomes in such a scenario can only be two. Therefore, the finite value is 2.

8.2.1 Definition: Let X be a discrete random taking values 1, 2, ……, n. Then X is said to follow uniform discrete uniform distribution if its p.m.f is given by

$$P(X = x) = \frac{1}{n} \qquad x = 1, 2,,\ldots n$$

$$= 0 \qquad \text{otherwise}$$

'n' is called as the parameter of the distribution. Whenever, the parameter value is known, the distribution is known completely. The name is 'uniform' as it treats all the values of the variable 'uniformly'. It is applicable where all the values of the random variable are equally likely.

Some examples or the situation where it applied

1. Let X denote the number on the face of unbiased die, after it is rolled.

$$P(X = x) = \frac{1}{6} \quad ; x = 1, 2, 3,4,5,6$$

$$= 0 \quad ; \text{otherwise}$$

2. A machine generates a digit randomly from 0 to 9

$$P(X = x) = \frac{1}{10} \quad ; x = 0, 1, 2, \ldots 9$$

$$= 0 \quad ; \text{otherwise}$$

8.2.2Mean and Variance of Uniform Distribution

Let X is said to follow Uniform discrete uniform distribution and its p.m.f is given by

$$P(X = x) = \frac{1}{n} \quad x = 1, 2,,\ldots n$$

$$= 0 \quad \text{otherwise}$$

Mean = E(X) = $\sum_{i=1}^{n} x_i * P(x_i) = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{n(n+1)}{2n} = \frac{n+1}{2}$

Variance *(X)* = $E(X^2) - [E(X)]^2$

$E(X^2) = \sum_{i=1}^{n} x^2{}_i * P(x_i) = \frac{1}{n}\sum_{i=1}^{n} x^2{}_i = \frac{(n+1)(2n+1)}{6}$

Variance *(X)* = $E(X^2) - [E(X)]^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}$

Standard Deviation (S.D.) (X)= $\sigma = \sqrt{Var\ (X)} = \sqrt{\frac{n^2-1}{12}}$

8.2.3 Applications of Uniform Distribution

Example1: Find the variance and standard deviation of X, where X is the square of the score shown on a fair die.

Solutions: Let X is a random variable which shows the square of the score on a fair die.

X = { 1, 4, 9,16, 25, 36 } each on have the probability = $\frac{1}{6}$

Mean = E(X) = $\sum_{i=1}^{n} x_i * P(x_i) = \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) = \frac{91}{6}$

Variance *(X)* = $E(X^2) - [E(X)]^2$

$E(X^2) = \sum_{i=1}^{n} x^2{}_i * P(x_i) = \frac{1}{6}(1 + 16 + 81 + 256 + 625 + 1296) = \frac{2275}{6}$

Variance *(X)* = $E(X^2) - [E(X)]^2 = \frac{2275}{6} - \left[\frac{91}{6} * \frac{91}{6}\right] = \frac{5396}{36}$

Standard Deviation (S.D.) (X)= σ = $\sqrt{Var\ (X)} = \sqrt{\frac{5396}{36}}$

8.3 Bernoulli Distribution

A trial is performed of an experiment whose outcome can be classified as either success or failure. The probability of success is p ($0 \le p \le 1$) and probability of failure is (1-p). A random variable X which takes two values 0 and 1 with probabilities 'q' and 'p' i.e. P(x=1) = p and P(x=0) = q, p + q = 1 (i.e q = 1 –p), is called a Bernoulli variate and is said to be a Bernoulli Distribution, where p and q takes the probabilities for success and failure respectively. It is discovered by Swiss Mathematician James or Jacques Bernoulli (1654-1705). It is applied whenever the experiment results in only two outcomes. One is success and other is failure. Such experiment is called as Bernoulli Trial.

8.3.1 Definition: Let X be a discrete random taking values either success (1/ True / p) or failure (0 / False / q). Then X is said to follow Bernoulli discrete distribution if its p.m.f is given by

$$P(X = x) = p^x q^{1-x} \qquad x = 0, 1$$

$$= 0 \qquad \text{otherwise}$$

Note: $0 \le p \le 1$, p + q = 1

This distribution is known as Bernoulli distribution with parameter 'p'

8.3.2 Mean and Variance of Bernoulli Distribution

Let X follows Bernoulli Distribution with parameter 'p'. Therefore its p.m.f. is given by

$$P(X = x) = p^x q^{1-x} \qquad x = 0, 1$$

$$= 0 \qquad \text{otherwise}$$

Note: $0 \le p \le 1$, $p + q = 1$

Mean = $E(X) = \sum_{i=0}^{1} x_i * P(x_i) = \sum_{i=0}^{1} x_i\, p^x q^{1-x}$

Substitute the value of $x = 0$, and $x = 1$, we get

$E(X) = (0 \times p^0 \times q^{1-0}) + (1 \times p^1 \times q^{1-1}) = p$

Similarly, $E(X^2) = \sum_{i=0}^{1} x_i^2 * P(x_i) = \sum_{i=0}^{1} x_i^2 p^x q^{1-x}$

Substitute the value of $x = 0$, and $x = 1$, we get

$E(X^2) = p$

Variance $(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p) = pq$ …… $(p + q = 1)$

Standard Deviation (S.D.) $(X) = \sigma = \sqrt{Var\,(X)} = \sqrt{pq}$

Note: if $p = q = \frac{1}{2}$ the Bernoulli distribution is reduced to a Discrete Uniform Distribution as

$$P(X = x) = \frac{1}{2} \qquad x = 0,1$$

$$= 0 \qquad \text{otherwise}$$

8.3.3 Applications of Bernoulli Distribution

Examples of Bernoulli' s Trails are:

1) Toss of a coin (head or tail)

2) Throw of a die (even or odd number)

3) Performance of a student in an examination (pass or fail)

4) Sex of a new born child is recorded in hospital, Male = 1, Female = 0

5) Items are classified as 'defective=0' and 'non-defective = 1'.

8.3.4 Distribution of Sum of independent and identically distributed Bernoulli Random variables

Let Y_i, i= 1,2,…n be 'n' independent Bernoulli random variables with parameter 'p' ('p' for success and 'q' for failure p + q = 1)

i.e. P [Y_i = 1] = p and P [Y_i = 0] = q, for i = 1,2,…n.

Now lets define, X which count the number of '1's (Successes) in 'n' independent Bernoulli trials,

$X = \sum_{i=1}^{n} Y_i$

In order to derive probability of '*x*' successes in 'n' trials i.e. P [X = *x*]

Consider a particular sequence of '*x'* successes and remaining (n – *x)* failures as

1 0 1 1 1 0 1 0 …… 1 0

Here '1' (Success = p) occurs '*x*' times and '0' (Failure = q) occurs (n – *x*) times.

Due to independence, probability of such a sequence is given as follows:

$$\underbrace{p\ p\ p \ldots . p}_{x\ times} \qquad \underbrace{q\ q\ q\ q \ldots \ldots \ldots . q}_{(n-x)times} = p^x q^{(n-x)}$$

However, the successes (1's) can occupy any *'x'* places out of ' n' places in a sequence in $\binom{n}{x}$ ways, therefore

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \quad x = 0, 1, 2 \ldots .n$$

$$= 0 \qquad \text{otherwise}$$

Note: $0 \le p \le 1$, p + q = 1

This gives us a famous distribution called as **' Binomial Distribution'**

8.4 Binomial Distribution

This distribution is very useful in day to day life. A binomial random variable counts number of successes when 'n' Bernoulli trials are performed. A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment, and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e., n = 1, the binomial distribution is a Bernoulli distribution.

Binomial distribution is denoted by $X \rightarrow B(n, p)$

Bernoulli distribution is just a binomial distribution with *n = 1*
i.e. parameters (1, *p)*.

8.4.1 Definition:

A discrete random variable X taking values 0, 1, 2,…….n is said to follow a binomial distribution with parameters 'n' and 'p' if its p.m.f. is given by

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \quad x = 0, 1, 2\ldots..n$$

$$= 0 \qquad \text{otherwise}$$

Note: $0 \le p \le 1$, $p + q = 1$

Remark:

1) The probabilities are term in the binomial expansion of $(p + q)^n$, hence the name ' Binomial Distribution ' is given

2) $\sum_{x=0}^{n} P(x) = \sum_{x=0}^{n} = \binom{n}{x} p^x q^{n-x} = (p + q)^n = 1$

3) The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one

8.4.2 Mean and Variance of Binomial Distribution

Let X follows Binomial Distribution with parameters 'n' and 'p' if its p.m.f. is given by

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \quad x = 0, 1, 2\ldots..n$$

$$= 0 \qquad \text{otherwise}$$

Note: $0 \le p \le 1$, $p + q = 1$

Mean = $E(X) = \sum_{i=0}^{1} x_i * P(x_i) = \sum_{i=0}^{1} x_i \binom{n}{x} p^x q^{n-x}$

Mean = $E(X) = \sum_{i=0}^{n} x_i * P(x_i)$

$= \sum_{i=0}^{n} x_i \binom{n}{x} p^x q^{n-x}$

$= \sum_{i=0}^{n} x_i \cdot \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$

Substitute the value of $x = 0$, we get first term as '0'

$= \sum_{i=1}^{n} \cdot \frac{(n-1)!}{(x-1)!\,(n-x)!} p^x q^{n-x}$

$= np \sum_{i=1}^{n} . \frac{(n-1)!}{(x-1)!\,(n-1-(x-1))!}\ p^{x-1}q^{n-x}$

$= np \sum_{i=1}^{n} . \binom{n-1}{x-1} p^{x-1}q^{n-1-(x-1)}$

$= np\,(p+q)^{n-1}$ ----- Using Binomial Expansion

$= np \ldots\ldots\ldots\ldots\ldots\ldots (p+q=1)$

Mean = E (X) = $\mu = np$ ----------(1)

$E(X^2) = E[X(X-1)] + E[X]$ ----------(2)

$E[X(X-1)] = \sum_{i=0}^{n} x_i (x_i - 1) \binom{n}{x} p^x q^{n-x}$

$= \sum_{i=0}^{n} \frac{x(x-1)n!}{x!\,(n-x)!}\ p^x q^{n-x}$

$= n(n-1)\,p^2 \sum_{i=2}^{n} . \binom{n-2}{x-2} p^{x-2}q^{n-2-(x-2)}$

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

$= n(n-1)\,p^2$----------(3)

Using (1), (2) and (3)

$E(X^2) = n(n-1)\,p^2 + np$ ----------(4)

Variance *(X)* = $E(X^2) - [E(X)]^2$

Variance *(X)* = $[n(n-1)\,p^2 + np] - (np)^2 = npq$

Standard Deviation (S.D.) (X)= $\sigma = \sqrt{Var(X)} = \sqrt{npq}$

NOTE: Binomial Theorem (Binomial Expansion) it states that, where n is a positive integer:

$(a+b)^n = a^n + ({}^nC_1)a^{n-1}b + ({}^nC_2)a^{n-2}b^2 + \ldots + ({}^nC_{n-1})ab^{n-1} + b^n$

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

8.4.3 Applications of Binomial Distribution

We get the Binomial distribution under the following experimental conditions.

1) The number of trials ' n' is finite. (not very large)

2) The trials are independent of each other.

3) The probability of success in any trial‘ p’ is constant for each trial.

4) Each trial (random experiment) must result in a success or a failure (Bernoulli trial).

Following are some of the real life examples of Binomial Distribution

1. Number of defective items in a lot of n items produced by a machine
2. Number of mail births out of ‘n’ births in a hospital
3. Number of correct answers in a multiple choice test.
4. Number of seeds germinated in a row of ‘n’ planted seeds
5. Number of rainy days in a month
6. Number of re-captured fish in a sample of ‘n’ fishes.
7. Number of missiles hitting the targets out of ‘n’ fired.

In all above situations, ‘p’ is the probability of success is assumed to be constant.

Example 1:

Comment on the following: “ The mean of a binomial distribution is 5 and its variance is 9”

Solution:

The parameters of the binomial distribution are n and p

We have mean $\Rightarrow$ np = 5

Variance $\Rightarrow$ npq = 9

$\therefore q = \frac{\text{npq}}{\text{np}} = \frac{9}{5} > 1$

Which is not admissible since q cannot exceed unity. (p + q = 1) Hence the given statement is wrong.

Example 2:

Eight coins are tossed simultaneously. Find the probability of getting atleast six heads.

Solution:

Here number of trials, n = 8, p denotes the probability of getting a head.

P = 1 / 2 and q = 1 – p = 1 / 2

If the random variable X denotes the number of heads, then the probability of a success in n trials is given by

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \qquad x = 0, 1, 2\ldots..n$$

$$= \binom{8}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{8-x} = \binom{8}{x}\left(\frac{1}{2}\right)^8 = \frac{1}{2^8}\binom{8}{x}$$

Note: $0 \le p \le 1$, $p + q = 1$

Probability of getting at least 6 heads is given by

$P(X \ge 6) = P(X = 6) + P(X = 7) + P(X = 8)$

$$= \frac{1}{2^8}\binom{8}{6} + \frac{1}{2^8}\binom{8}{7} + \frac{1}{2^8}\binom{8}{8}$$

$$= \frac{1}{2^8}\left[\binom{8}{6} + \binom{8}{7} + \binom{8}{8}\right]$$

$$= \frac{1}{256}[\,28 + 8 + 1\,] = \frac{37}{256}$$

Example 3:

Ten coins are tossed simultaneously. Find the probability of getting (i) at least seven heads (ii) exactly seven heads (iii) at most seven heads

Solution:

Here number of trials, n = 10, p denotes the probability of getting a head.

P = 1 / 2 and q = 1 – p = 1 / 2

If the random variable X denotes the number of heads, then the probability of a success in n trials is given by

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \qquad x = 0, 1, 2\ldots..n$$

$$= \binom{10}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{10-x} = \binom{10}{x}\left(\frac{1}{2}\right)^{10} = \frac{1}{2^{10}}\binom{10}{x}$$

Note: $0 \le p \le 1$, $p + q = 1$

i) Probability of getting at least 7 heads is given by

$$P(X \geq 7) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$$

$$= \frac{1}{2^{10}}\binom{10}{7} + \frac{1}{2^{10}}\binom{10}{8} + \frac{1}{2^{10}}\binom{10}{9} + \frac{1}{2^{10}}\binom{10}{10}$$

$$= \frac{1}{2^{10}}\left[\binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right]$$

$$= \frac{1}{1024}[\,120 + 45 + 10 + 1\,] = \frac{176}{1024}$$

ii) Probability of getting exactly 7 heads is given by .

$$P(X = 7) = \frac{1}{2^{10}}\binom{10}{7} = \frac{120}{1024}$$

iii) Probability of getting at most 7 heads is given by .

$$P(X \leq 7) = 1 - P(X > 7)$$

$$= 1 - \{\, P(X = 8) + P(X = 9) + P(X = 10)\,\}$$

$$= 1 - \frac{1}{2^{10}}\left[\binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right]$$

$$= 1 - \frac{1}{1024}[\,45 + 10 + 1\,] = 1 - \frac{56}{1024} = \frac{968}{1024}$$

Example 4:

20 wrist watches in a box of 100 are defective. If 10 watches are selected at random, find the probability that (i) 10 are defective (ii) 10 are good (iii) at least one watch is defective (iv)at most 3 are defective.

Solution:

20 out of 100 wrist watches are defective, so Probability of defective wrist watch p = 20 /100

$$p = \frac{20}{100} = \frac{1}{5} \quad \therefore q = 1 - p = 1 - \frac{20}{100} = \frac{80}{100} = \frac{4}{5}$$

Since 10 watches are selected at random, n = 10

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \qquad x = 0, 1, 2\ldots..n$$

$$= \binom{10}{x}\left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{10-x}$$

i) Probability of selecting 10 defective watches

$$P(X = 10) = \binom{10}{10}\left(\frac{1}{5}\right)^{10}\left(\frac{4}{5}\right)^{10-10} = 1.\frac{1}{5^{10}}.1 = \frac{1}{5^{10}}$$

ii) Probability of selecting 10 good watches (i.e. no defective)

$$P(X = 0) = \binom{10}{0}\left(\frac{1}{5}\right)^{0}\left(\frac{4}{5}\right)^{10-0} = 1.1.\frac{4^{10}}{5^{10}} = \left(\frac{4}{5}\right)^{10}$$

iii) Probability of selecting at least one defective watch

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0)$$

$$= 1 - \binom{10}{0}\left(\frac{1}{5}\right)^{0}\left(\frac{4}{5}\right)^{10-0} = 1 - \left(\frac{4}{5}\right)^{10}$$

iv) Probability of selecting at most 3 defective watches

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$= \binom{10}{0}\left(\frac{1}{5}\right)^{0}\left(\frac{4}{5}\right)^{10} + \binom{10}{1}\left(\frac{1}{5}\right)^{1}\left(\frac{4}{5}\right)^{9} + \binom{10}{2}\left(\frac{1}{5}\right)^{2}\left(\frac{4}{5}\right)^{8} + \binom{10}{3}\left(\frac{1}{5}\right)^{3}\left(\frac{4}{5}\right)^{7}$$

$$= 1.1.\left(\frac{4}{5}\right)^{10} + 10.\frac{1}{5}\frac{4^9}{5^9} + 45.\frac{1}{5^2}\frac{4^8}{5^8} + 120.\frac{1}{5^3}\frac{4^7}{5^7}$$

$$= 0.859 \text{ (Approx.)}$$

Example 5:

With the usual notation find p for binomial random variable X if n = 6 and [9*P(X = 4) = P(X = 2)]

Solution:

The probability mass function of binomial random variable X is given by

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \qquad x = 0, 1, 2\ldots..n$$

Note: $0 \leq p \leq 1$, $p + q = 1$

Here n = 6,

$$P(X = x) = \binom{6}{x}(p)^x(q)^{6-x}$$

$$P(X = 4) = \binom{6}{4}(p)^4(q)^{6-4} = \binom{6}{4}(p)^4(q)^2$$

$$P(X = 2) = \binom{6}{2}(p)^2(q)^{6-2} = \binom{6}{2}(p)^2(q)^4$$

Given that, $9*P(X = 4) = P(X = 2)$

$$9 * \binom{6}{4}(p)^4(q)^2 = \binom{6}{2}(p)^2(q)^4$$

$$9 * 15* (p)^4(q)^2 = 15 * (p)^2(q)^4$$

$$9 (p)^4(q)^2 = (p)^2(q)^4$$

$$9 p^2 = q^2$$

Taking positive square root on both sides we get,

$$3p = q$$

$$3p = 1- p$$

$$4p = 1 \ \therefore p = \frac{1}{4} = 0.25$$

8.5 Poisson Distribution

Poisson distribution was discovered by a French Mathematician-cum-Physicist Simeon Denis Poisson in 1837. Poisson distribution is also a discrete distribution. He derived it as a limiting case of Binomial distribution. For n-trials the binomial distribution is $(p + q)^n$; the probability of x successes is given by $P(X=x)= \sum_{i=0}^{n} x_i \binom{n}{x} p^x q^{n-x}$ If the number of trials n is very large and the probability of success ' p' is very small so that the product np = m is non – negative and finite.

8.5.1 Definition:

A discrete random variable X, taking on one of the countable infinite values 0, 1, 2,……. is said to follow a Binomial distribution with parameters 'λ' or 'm', if for some m > 0, its p.m.f. is given by

$$P(X = x) = \frac{e^{-m}m^x}{x!} \qquad x = 0, 1, 2\ldots..$$

$$m > 0$$

$$= 0 \qquad \text{otherwise}$$

Note: $e^{-m} \geq 0$, $AND\ m^x \geq 0$, $x! \geq 0$,

Hence, $\frac{e^{-m}m^x}{x!} \geq 0$

Note: $e^m = \frac{m^0}{0!} + \frac{m^1}{1!} + \frac{m^2}{2!} - - - - - - = \sum_{x=0}^{\infty} \frac{m^x}{x!}$

$e = \frac{1^0}{0!} + \frac{1^1}{1!} + \frac{2^2}{2!} - - - - - - = 2.71828$, 0!=1, 1!=1

Since, $e^{-m} \geq 0 \ \therefore P(X) \geq 0$ for all x and

$$\sum P(x) = \sum_{x=0}^{\infty} \frac{e^{-m}m^x}{x!} = e^{-m} \ , \sum_{x=0}^{\infty} \frac{m^x}{x!} = e^{-m}.e^m = 1$$

It is denoted by X → P(m) or X → P(λ)

Since number of trials is very large and the probability of success p is very small, it is clear that the event is a rare event. Therefore Poisson distribution relates to rare events.

8.5.2 Mean and Variance of Poisson Distribution

Let X is a Poisson random variable with parameter 'm' (or 'λ'), if its p.m.f. is given as

$$P(X = x) = \frac{e^{-m}m^x}{x!} \qquad x = 0, 1, 2\ldots.. \quad m > 0$$

$$= 0 \qquad \text{otherwise}$$

Note: $e^{-m} \geq 0,\ AND\ m^x \geq 0, x! \geq 0,$

Hence, $\frac{e^{-m}m^x}{x!} \geq 0$

∴ Mean = E(X) = $\sum_{i=0}^{\infty} x_i\, P(x_i)$

$$= \sum_{i=0}^{\infty} x_i \frac{e^{-m}m^x}{x!}$$

The term corresponding to $x = 0$ is zero.

$$\therefore \quad = \sum_{i=1}^{\infty} m \frac{e^{-m}\, m^{x-1}}{(x-1)!} = me^{-m} \sum_{i=1}^{\infty} \frac{e^{-m}\, m^{x-1}}{(x-1)!} = me^{-m} \cdot e^{m} = m$$

∴ Mean = E(X) = μ = m

But, $E(X^2) = E[X(X-1)] + E[X]$

$$E[X(X-1)] = \sum_{i=0}^{\infty} x_i\,(x_i - 1)\ P(x_i) = \sum_{i=0}^{\infty} x_i\,(x_i - 1)\ \frac{e^{-m}m^x}{x!}$$

$$= m^2e^{-m} \sum_{i=2}^{\infty} \frac{m^{x-2}}{(x-2)!} = m^2e^{-m} \cdot e^{m} = m^2$$

$E[X(X-1)] = m^2$

$E(X^2) = E[X(X-1)] + E[X] = m^2 + m$

Variance *(X)* $= E(X^2) - [E(X)]^2$

Variance *(X)* $= (m^2 + m) - (m)^2 = m$

Standard Deviation (S.D.) (X)= σ = $\sqrt{Var(X)} = \sqrt{m}$

Thus the mean and variance of Poisson distribution are equal and each is equal to the parameter of distribution ($'m'$ or $'\lambda'$)

8.5.3 Applications of Poisson Distribution

Some examples of Poisson variates are :

1) The number of blinds born in a town in a particular year.
2) Number of mistakes committed in a typed page.
3) The number of students scoring very high marks in all subjects.
4) The number of plane accidents in a particular week.
5) The number of defective screws in a box of 100, manufactured by a reputed company.
6) Number of accidents on the express way in one day.
7) Number of earthquakes occurring in one year in a particular seismic zone.
8) Number of suicides reported in a particular day.
9) Number of deaths of policy holders in one year.

Conditions:

Poisson distribution is the limiting case of binomial distribution under the following conditions:

1. The number of trials n is indefinitely large i.e., $n \to \infty$
2. The probability of success ' p' for each trial is very small; i.e., $p \to 0$
3. np = m (say) is finite, $m > 0$

Characteristics of Poisson Distribution:

Following are the characteristics of Poisson distribution

1. Discrete distribution: Poisson distribution is a discrete distribution like Binomial distribution, where the random variable assume as a countably infinite number of values 0,1,2 ….
2. The values of p and q: It is applied in situation where the probability of success p of an event is very small and that of failure q is very high almost equal to 1 and n is very large.
3. The parameter: The parameter of the Poisson distribution is m. If the value of m is known, all the probabilities of the Poisson distribution can be ascertained.

4. Values of Constant: Mean = m = variance; so that standard deviation = $\sqrt{m}$ Poisson distribution may have either one or two modes.

5. Additive Property: If X_1 and X_2 are two independent Poisson distribution variables with parameter m_1 and m_2 respectively. Then $(X_1 + X_2)$ also follows the Poisson distribution with parameter $(m_1 + m_2)$ i.e. $(X_1 + X_2) \rightarrow P(m_1 + m_2)$

6. As an approximation to binomial distribution: Poisson distribution can be taken as a limiting form of Binomial distribution when n is large and p is very small in such a way that product np = m remains constant.

7. Assumptions: The Poisson distribution is based on the following assumptions.

   i) The occurrence or non- occurrence of an event does not influence the occurrence or non-occurrence of any other event.

   ii) The probability of success for a short time interval or a small region of space is proportional to the length of the time interval or space as the case may be.

   iii) The probability of the happening of more than one event is a very small interval is negligible.

Example 1:

Suppose on an average 1 house in 1000 in a certain district has a fire during a year. If there are 2000 houses in that district, what is the probability that exactly 5 houses will have a fire during the year? [given that $e^{-2} = 0.13534$]

Solution:

Mean = np , n = 2000, p = $\frac{1}{1000}$

m = np = $2000 * \frac{1}{1000} = 2$

∴ m = 2, now According to Poisson distribution

$$P(X = x) = \frac{e^{-m} m^x}{x!} \qquad x = 0, 1, 2\ldots.. \quad m > 0$$

$$\therefore \mathbf{P(X = 5)} = \frac{e^{-m} m^x}{x!} = \frac{e^{-2} 2^5}{5!}$$

$$\mathbf{P(X = 5)} = \frac{(0.13534) * 32}{120} = 0.36$$

Example 2:

In a Poisson distribution 3P(X=2) = P(X=4) Find the parameter ' m' .

Solution:

$$P(X = x) = \frac{e^{-m}m^{x}}{x!} \qquad x = 0, 1, 2\ldots.. \quad m > 0$$

Given, 3P(X=2) = P(X=4)

$$\therefore \; 3\,\frac{e^{-m}m^{2}}{2!} = \frac{e^{-m}m^{4}}{4!}$$

$$\therefore \; m^{2} = \frac{3.4!}{2!} = 36$$

$$\therefore \; m = \pm 6$$

Since mean is always positive $\therefore$ m = 6

Example 3:

If 2% of electric bulbs manufactured by a certain company are defective. Find the probability that in a sample of 200 bulbs i) less than 2 bulbs ii) more than 3 bulbs are defective.[$e^{-4} = 0.0183$]

Solution :

The probability of a defective bulb = $p = 2 / 100 = 0.02$

Given that n=200 since p is small and n is large, we use Poisson Distribution, mean m = np

m = np = 200 * 0.02 = 4

Now, Poisson Probability Function

$$P(X = x) = \frac{e^{-m}m^{x}}{x!} \qquad x = 0, 1, 2\ldots.. \quad m > 0$$

i) Probability of less than 2 bulbs are defective

P (X < 2) = P (X=0) + P (X = 1)

$$= \frac{e^{-4}4^{0}}{0!} + \frac{e^{-4}4^{1}}{1!} = e^{-4}(1 + 4) = e^{-4} * 5 = 0.0138 * 5 = 0.0915$$

ii) Probability of getting more than 3 defective bulbs

$P(X > 3) = 1 - P(X \leq 3)$

$= 1 - \{ P(X=0) + P(X=1) + P(X=2) + P(X=3) \}$

$= 1 - e^{-4} \left(1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!}\right)$

$= 1 - 0.0183*(1+4+8+10.67)$

$= 0.567$

Example 4:

In a company previous record show that on an average 3 workers are absent without leave per shift. Find the probability that in a shift

i) Exactly 2 workers are absent

ii) More than 4 workers will be absent

iii) At most 3 workers will be absent

Solution :

This is a case of Poisson distribution with parameter 'm=3'

i) $P(X = 2) = \frac{e^{-3}3^2}{3!} = 0.2241$

ii) $P(X > 4) = 1 - P(X \leq 4)$

$= 1 - \{ P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) \}$

$= 1 - \{ \frac{e^{-3}3^0}{0!} + \frac{e^{-3}3^1}{1!} + \frac{e^{-3}3^2}{2!} + \frac{e^{-3}3^3}{3!} + \frac{e^{-3}3^4}{4!} \} = 0.1845$

iii) $P(X \geq 3) = 1 - P(X \leq 3)$

$= 1 - \{ P(X=0) + P(X=1) + P(X=2) + P(X=3) \}$

$= 1 - \{ \frac{e^{-3}3^0}{0!} + \frac{e^{-3}3^1}{1!} + \frac{e^{-3}3^2}{2!} + \frac{e^{-3}3^3}{3!} \} = 0.5767$

Example 5:

Number of accidents on Pune Mumbai express way each day is a Poisson random variable with average of three accidents per day. What is the probability that no accident will occur today?

Solution :

This is a case of Poisson distribution with m = 3,

$P(X = 0) = \frac{e^{-m}3^0}{0!} = e^{-3} = 0.0498$

Example 6:

Number of errors on a single page has Poisson distribution with average number of errors one per page . Calculate probability that there is at least one error on a page.

Solution :

This is a case of Poisson distribution with m = 1,

$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{e^{-1}1^0}{0!} = 1 - e^{-1} = 0.0632$

8.6 Summary

In this chapter, Discrete distribution, its types Uniform, Bernoulli, Binomial and Poisson with its mean , variance and its application is discussed.

Distribution	**Definition**	**Mean E(X)**	**Variance (X)**
Uniform	$P(X = x) = \frac{1}{n}$ $x = 1, 2,,\ldots n$ $= 0$ otherwise	$\frac{n+1}{2}$	$\sqrt{\frac{n^2-1}{12}}$
Bernoulli	$P(X = x) = p^x q^{1-x}$ $x = 0, 1$ $= 0$ otherwise Note: $0 \leq p \leq 1$, $p + q = 1$	p	pq
Binomial	$P(X = x) = \binom{n}{x} p^x q^{n-x}$ $x = 0, 1,2,\ldots n$ $= 0$ otherwise Note: $0 \leq p \leq 1$, $p + q = 1$	$\mu = np$	npq
Poisson	$P(X = x) = \frac{e^{-m}m^x}{x!}$ $x = 0, 1, 2\ldots..$ $m > 0$ $= 0$ Otherwise	$\mu = m$	m

8.7 Reference

Fundamentals of Mathematical Statistics S. C. Gupta, V. K. Kapoor

UNIT 4

9

FITTING OF CURVES

Unit Structure

9.0 Objectives

From this chapter student should learn analysis of data using various methods. Methods involve moving average method and least square method seasonal fluctuations can be studied by business for casting method.

9.1 Introduction

Every business venture needs to know their performance in the past and with the help of some predictions based on that, would like to decide their strategy for the present By studying the past behavior of the characteristics, the nature of variation in the value can be determined. The values in the past can be compared with the present values of comparisons at different places during formulation of future plan and policies. This is applicable to economic policy makers, meteorological department, social scientists, political analysis. Forecasting thus is an important tool in Statistical analysis. The statistical data, particularly in the field of social science, are dynamic in nature. Agricultural and Industrial production increase every year or due to improved medical facilities, there is decline in the death rate over a period of time. There is increase in sales and exports of various products over a period of years. Thus, a distinct change (either increasing or decreasing) can be observed in the value of time-series.

A time series is a sequence of value of a phenomenon arranged in order of their occurrence. Mathematically it can expressed as a function, namely $y = f(t)$ where t represents time and y represents the corresponding values. That is, the value y_1,

y_2, y_3 …… of a phenomenon with respect to time periods t_1, t_2, t_3 …. Form a Time Series.

Forecasting techniques facilities prediction on the basic of a data available from the past. This data from the past is called a time series. A set of observations, of a variable, taken at a regular (fixed and equal) interval of time is called time series. A time series is a bivariate data, with time as the independent variable and the other is the variable under consideration. There are various forecasting method for time series which enable us to study the variation or trends and estimate the same for the future.

9.2 Importance of Time Series Analysis

The analysis of the data in the time series using various forecasting model is called as time analysis. The importance of time series analysis is due to the following reasons:

- *Understanding the past behavior*
- *Planning the future action*
- *Comparative study*

9.3 Components of Time Series

The fluctuation in a time series are due to one or more of the following factors which are called "components" of time series.

(a) Secular Trend :

The general tendency of the data, either to increase, to decrease or to remain constant is called Secular Trend. It is smooth, long term movement of the data. The changes in the values are gradual and continuous. An increasing demand for luxury items like refrigerators or colour T.V. sets reflect increasing trend. The production of steel, cement, vehicles shows a rising trend. On the other hand, decreasing in import of food grains is an example of decreasing trend. The nature of the trend may be linear or curvilinear, in practice, curvilinear trend is more common.

Trend in due to long term tendency. Hence it can be evaluation if the time series is a available over a long duration.

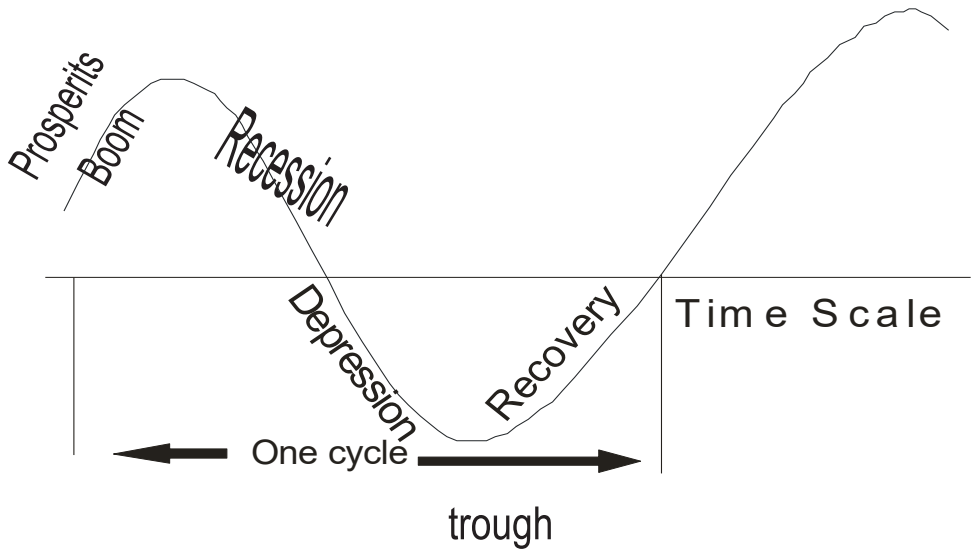
(b) Seasonal Variation :

The regular, seasonal change in the time series are called "Seasonal Variation". It is observed that the demand for umbrellas, raincoats reaches a peak during monsoon or the advertisement of cold drinks, ice creams get a boom in summer. The demand for greeting cards, sweets, increase during festival like Diwali, Christmas. In March, there is maximum withdrawal of bank deposits for adjustment of income-tax payment, so also variation tax-saving schemes shoot up during this period.

The causes, for these seasonal fluctuations, are thus change in weather conditions, the traditions and customs of people etc. The seasonal component is measured to isolate these change from the trend component and to study their effect, so that, in any business, future production can be planned accordingly and necessary adjustments for seasonal change can be made

(c) Cyclical Variation :

These are changes in time series, occurring over a period which is more than a year. They are recurring and periodic in nature. The period may not be uniform. These fluctuations are due to changes in a business cycle. There are four important phases of any business activity viz. prosperity, recession, depression and recovery. During prosperity, the business flourishes and the profit reaches a maximum level. Thereafter, in recession, the profit decreases, reaching a minimum level during depression. After some time period, the business again recovers (recovery) and it is followed by period of prosperity. The variation in the time series due to these phases in a business cycle are called "Cyclical Variation:.



The knowledge of cyclic variations is important for a businessman to plan his activity or design his policy for the phase of recession or depression. But one should know that the factors affecting the cyclical variations are quite irregular, difficult to identify and measure. The cyclical variation are denoted by

(d) Irregular Variation :

The changes in the time series which can not be predicated and are erratic in nature are called "Irregular Variation". Usually, these are short term changes having signification effect on the time series during that time interval. These are caused by unforeseen event like wars, floods, strikes, political charges, etc. During Iran-Iraq war or recent Russian revolution, prices of petrol and petroleum product soared very high. In recent budget, control on capital issued was suddenly removed. As an effect, the all Indian-Index of share market shooted very high, creates all time records. If the effect of other components of the time series is eliminated, the remaining variation are called "Irregular or Random Variations". No forecast of these change can be made as they do not reflect any fixed pattern.

MODELS FOR ANALYSIS OF TIME SERIES

The purpose of studying time series is to estimate or forecast the value of the variable. As there are four components of the time series, these are to be studied separately. There are two types of models which are used to express the relationship of the components of the time series. They are additive model and multiplicative model.

O = Original Time Series

T = Secular Trend

S = Seasonal Variations

C = Cyclical Variations and

I = Irregular Variations

In Additive model, it is assumed that the effect of the individual components can be added to get resultant value of the time series, that is the components are independent of one another. The model can be expressed as

$$O = T + S + C + I$$

In multiplicative model, it is assumed that the multiplication of the individual effect of the components result in the time series, that is, the components are due to different causes but they are not necessarily independent, so that changes in any one of them can affect the other components. This model is more commonly used. It is expressed as

$$O = T \times S \times C \times I$$

If we want to estimate the value in time series, we have to first estimate the four components and them combine them to estimate the value of the time series. The irregular variations can be found. However, we will restrict ourselves, to discuss method of estimating the first components, namely Secular Trend.

9.4 Methods to Find Trend

There are various method to find the trend. The major methods are as mentioned below:

I. Free Hand Curve.

II. Method of Semi – Averages.

III. Method of Moving Averages.

IV. Method of Least Squares.

we will study only the method of moving average and least squares.

9.4.1.Method of Moving Averages

This is a simple method in which we take the arithmetic average of the given times series over a certain period of time. These average move over period and are hence called as moving averages. The time interval for the average is taken as 3 years, 4 years or 5 years and so on. The average are thus called as 3 yearly, 4 yearly and 5 yearly moving average. The moving averages are useful in smoothing the fluctuations caused to the variable. Obviously larger the time interval of the average more is the smoothing. We shall study the odd yearly (3 and 5) moving average first and then the 4 yearly moving average.

Odd Yearly Moving Average

In this method the total of the value in the time series is taken for the given time interval and is written in front of the middle value. The average so taken is also written in front of this middle value. This average is the trend for that middle year. The process is continued by replacing the first value with the next value in the time series and so on till the trend for the last middle value is calculated. Let us understand this with example:

Example 1:

Find 3 years moving averages and draw these on a graph paper. Also represent the original time series on the graph.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007
Production (in thousand unit)	12	15	20	18	25	32	30	40	44

Solution:

We calculate arithmetic mean of first three observations viz. 12, 15 and 20, then we delete 12 and consider the next one so that now, average of 15, 20 and 18 is calculated and so on. These averages are placed against the middle year of each group, viz. the year 2000, 2001 and so on. Note moving averages are not obtained for the year 1999 and2007.

Year	Production (in thousand unit)	3 Years Total	3yrly.Moving Average
1999	12		
2000	15	12 + 15 + 20 = 47	47 / 3 = 15.6
2001	20	15 + 20 + 18 = 53	53 / 3 = 17.6
2002	18	20 + 18 + 25 = 63	63 / 3 = 21.0
2003	25	18 + 25 + 32 = 75	75 / 3 = 25.0
2004	32	25 + 32 + 30 = 87	87 / 3 = 29.0
2005	30	32 + 30 + 40 = 102	102 / 3 = 34.0
2006	40	30 + 40 + 44 = 114	114 / 3 = 38.0
2007	44		

Example 2:

Find 5 yearly moving average for the following data.

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Sales (in lakhs of Rs.)	51	53	56	57	60	55	59	62	68	70

Solution:

We find the average of first five values, namely 51, 53, 56, 57 and 60. Then we omit the first value 51 and consider the average of next five values, that is, 53, 56, 57, 60 and 55. This process is continued till we get the average of the last five values 55, 59, 62, 68 and 70. The following table is prepared.

Year	Sales (in lakhs of Rs.)	5 Years Total	Moving Average (Total / 5)
1997	51	….	….
1998	53	….	….
1999	56	51 + 53 + 56 + 57 +60 = 277	55.4
2000	57	53 + 56 + 57 + 60 + 55 = 281	56.2
2001	60	56 + 57 + 60 + 55 + 59 = 287	57.4

Year	Sales (in lakhs of Rs.)	5 Years Total	Moving Average (Total / 5)
2002	55	57 + 60 + 55 + 59 + 62 = 293	58.6
2003	59	60 + 55 + 59 + 62 + 68 = 304	60.8
2004	62	55 + 59 + 62 + 68 + 70 = 314	62.8
2005	68	….	….
2006	70	….	….

Example 3:

Determine the trend of the following time series using 5 yearly moving averages.

Year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Exports in'000Rs	78	84	80	83	86	89	88	90	94	93	96

Solution: The time series is divided into overlapping groups of five years, their 5 yearly total and average are calculated as shown in the following table.

Year	Export (Y)	5 – yearly total (T)	5 – yearly moving average: (T/5)
1981	78		
1982	84		
1983	80	78+84+80+83+86 = 411	411 / 5 = 82.2
1984	83	84+80+83+86+89 = 422	422 / 5 = 84.4
1985	86	80+83+86+89+88 = 426	85.2
1986	89	83+86+89+88+90 = 436	87.2
1987	88	86+89+88+90+94 = 447	89.4
1988	90	89+88+90+94+93 = 454	90.8
1989	94	88+90+94+93+96 = 461	92.2
1990	93		
1991	96		

Observations:

I. In case of the 5 – yearly moving average, the total and average for the first two and the last two in the time series is not calculated. Thus, the moving average of the first two and the last two years in the series cannot be computed.

II. To find the 3 – yearly total (or 5 – yearly total) for a particular years, you can subtract the first value from the previous year's total, and add the next value so as to save your time!

Even yearly moving averages

In case of even yearly moving average the method is slightly different as here we cannot find the middle year of the four years in consideration. Here we find the total for the first four years and place it between the second and the third year value of the variable. These totals are again sunned into group of two, called as centered total and is placed between the two totals. The 4 – yearly moving average is found by dividing these centered totals by 8. Let us understand this method with an example

Example 4: Calculate the 4 yearly moving averages for the following data.

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999
Import in'000Rs	15	18	20	24	21	25	28	26	30

Ans: The table of calculation is show below. Student should leave one line blank after every to place the centered total in between two years.

Years	**(Import) Y**	**4 - Yearly Total**	**4 - Yearly Centered Total**	**4 - Yearly Moving Averages :**
1991	15	-	-	
1992	18	-	-	
1993	20	77	77 + 83 = 160	160/8 = 20
1994	24	83	83 + 90 = 173	173/8 = 21.6
1995	21	90	90 + 98 = 188	188/8 = 23.5
1996	25	98	98 + 100 = 198	198/8 = 24.8
1997	28	100	100 + 109 = 209	209/8 = 26.1
1998	26	109	-	-
1999	30	-	-	-

Example 5:

Find the moving average of length 4 for the following data. Represent the given data and the moving average on a graph paper.

Year	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Sales (in thousand unit)	60	69	81	86	78	93	102	107	100	109

Solution: We prepare the following table.

Year	Sale (in thousand unit)	4 Yearly Totals	Centred Total	Moving / Avg. Central = Total / 8
1998	60			
1999	69			
		60 + 69 + 81 + 86 = 296		
2000	81		296 + 314 = 610	76.25
		69 + 81 + 86 + 78 = 314		
2001	86		314 + 338 = 652	81.5
		81 + 86 + 78 + 93 = 338		
2002	78		338 + 359 = 697	87.125
		86 + 78 + 93 + 102 = 359		
2003	93		359 + 380 = 739	92.375
		78 + 93 + 102 + 107 = 380		
2004	102		380 + 402 = 782	97.75
		93 + 102 + 107 + 100 = 402		
2005	107		402 + 418 = 820	102.5
		102 + 107 + 100 + 109 = 418		
2006	100			
2007	109			

Note that 4 yearly total are written between the years 1999-2000, 2000-01, 2001-02 etc. and the central total are written against the years 2000, 2001, 2002 etc. so also the moving average are considered w.r.t. years; 2000, 2001 and so on. The moving averages are obtained by dividing the certain total by 8.

The graph of the given set of values and the moving averages against time representing the trend component are shown below. Note that the moving averages are not obtained for the years 1998, 1999, 2006 and 2007. (i.e. first and last two extreme years).

When the values in the time series are plotted, a rough idea about the type of trend whether linear or curvilinear can be obtained. Then, accordingly a linear or second degree equation can be fitted to the values. In this chapter, we will discuss linear trend only.

9.4.2. LEAST SQUARES METHOD:

Let $y = a + bx$ be the equation of the straight line trend where a, b are constant to be determined by solving the following normal equations,

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

where y represents the given time series.

We define x from years such that $\sum x = 0$. So substituting $\sum x = 0$ in the normal equation and simplifying, we get

$$b = \frac{\sum xy}{\sum x^2} \text{ and } a = \frac{\sum y}{n}$$

Using the given set of values of the time series, a, b can be calculated and the straight line trend can be determined as y = a + bx. This gives the minimum sum of squares line deviations between the original data and the estimated trend values. The method provides estimates of trend values for all the years. The method has mathematical basis and so element of personal bias is not introduced in the calculation. As it is based on all the values, if any values are added, all the calculations are to be done again.

Odd number of years in the time series

When the number of years in the given time series is add, for the middle year we assume the value of x = 0. For the years above the middle year the value given to x are …, -2, -1 while those after the middle year are values 1,2, … and so on.

Even number of years in the time series

When the number of years in the time series is even, then for the upper half the value of x are assumed as…., -5, -3, -1. For the lower half years, the values of x are assumed as 1, 3, 5, …. And so on.

Example 6:

Fit a straight line trend for the following data giving the annual profits (in lakhs of Rs.) of a company. Estimate the profit for the year 1999.

Years	1992	1993	1994	1995	1996	1997	1998
Profit	30	34	38	36	39	40	44

Solution: Let y = a + bx be the straight line trend.

The number of years is seven, which is sold. Thus, the value of x is taken as 0 for the middle years 1995, for upper three years as -3, -2, -1 and for lower three years as 1, 2, 3.

The table of computation is as shown below:

Years	**Profit (y)**	**x**	**xy**	**x^2**	**Trend Value: $Y_t = a + bx$**
1992	30	-3	-90	9	31.41
1993	34	-2	-68	4	33.37
1994	38	-1	-38	1	35.33
1995	36	0	0	0	37.29
1996	39	1	39	1	39.25
1997	40	2	80	4	41.21
1998	44	3	132	9	43.17
Total	$\sum y = 261$	$\sum x = 0$	$\sum xy = 55$	$\sum x^2 = 28$	

From the table : n = 7, $\sum xy = 55$, $\sum x^2 = 28$, $\sum y = 261$

There fore $b = \frac{\sum xy}{\sum x^2} = \frac{55}{28} = 1.96$ and $a = \frac{\sum y}{n} = \frac{261}{7} = 37.29$

Thus, the straight line trend is y = 37.29 + 1.96x.

The trend values in the table for the respective years are calculated by substituting the corresponding value of x in the above trend line equation.

For the trend value for 1992: x = -3:

$y_{1992} = 37.29 + 1.96\ (-3) = 37.29 - 5.88 = 31.41$

Similarly, all the remaining trend values are calculated.

(A short-cut method in case of odd number of years to find the remaining trend values once we calculate the first one, is to add the value of b to the first trend value to get the second trend value, then to the second trend value to get the third one and so on. This is because the difference in the values of x is 1.)

To estimate the profit for the years 1999 in the trend line equation, we substitute the prospective value of x, if the table was extended to 1999. i.e. we put x = 4, the next value after x = 3 for the year 1998.

$\therefore y_{1999} = 37.29 + 1.96\ (4) = 45.13$

There fore the estimated profit for the year 1999 is Rs. 45.13 lakhs.

Example 7:

Fit straight line trend by the method of lease squares for the following data representing production in thousand units. Plot the data and the trend line on a graph paper. Hence or otherwise estimate the trend for the years 2007.

Year	1999	2000	2001	2002	2003	2004	2005
Production (in thousand unit)	14	15	17	16	17	20	23

Solution:

Here, the total number of years is 7, an odd number. So we take the center as 1986 the middle-most year and define x as year 2002. The values of x will be -3, -2, -1, 0, 1, 2, 3.

Prepare the following table to calculate the required summations. Note that the trend values can be written in the table only after calculation of a and b.

Year	**Production (y)**	**x**	**x^2**	**x y**	**Trend Values**
1999	14	-3	9	-42	13.47
2000	15	-2	4	-30	14.79
2001	17	-1	1	-17	16.11
2002	16	0	0	0	17.43
2003	17	1	1	17	18.75
2004	20	2	4	40	20.07
2005	23	3	3	69	21.39
	122		28	37	

Here, n = 7, $\sum y = 122$, $\sum x^2 = 28$, $\sum x y = 37$

Now, a and b are calculated as follows:

$$a = \frac{\sum y}{n} = \frac{122}{7} = 17.4286 \approx 17.43$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{37}{28} = 1.3214 \approx 1.32$$

So, the equation is used to find trend values.

$y = a + b x$

i.e. $y = 17.43 + 1.32x$

The equation is used to find trend values.

For the year 1999, x = -3, substituting the value od x, we get,

y = 17.43 + 1.32 (-3) = 17.43 – 3.96 = 13.47

to find the remaining trend values we can make use of the property of a straight line that as all the values of x are equidistant with different of one unit (-3, -2, -1, --- and so on), the estimated trend value will also be equidistant with a difference of b unit.

In this case as b = 1.32, the remaining trend values for x = -2, -1, 0, --- etc. are obtained by adding b = 1.32 to the previous values. So, the trend values are 13.47, 14.79, 16.11, 17.43, 18.75, 20.07 and 21.39.

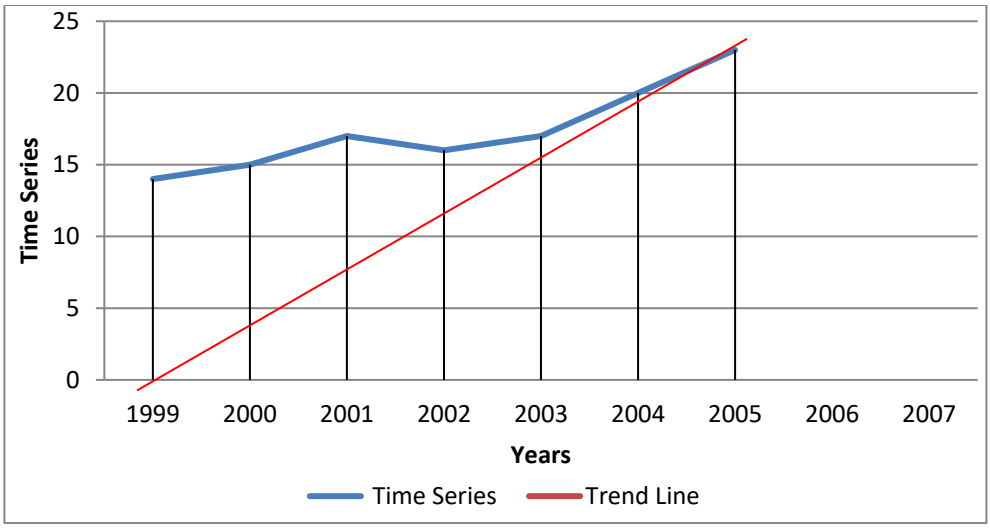
Now to estimate trend for the year 2007, x = 5, substituting in the equation

y = 17.43 + 1.32x

= 17.43 + 1.32 (5) = 24.03

So, the estimated trend value for the year 2007 is 24,030 unit.

For graph of time series, all points are plotted. But for the graph of trend line, any two trend values can be plotted and the line joining these points represents the straight line trend.



For the trend line, the trend values 17.43 and 21.39 for the years 2002 and 2005 are plotted and then a straight line joining these two points is drawn and is extended on both the sides.

The estimate of trend for the year 2007 can also be obtained from the graph by drawing a perpendicular for the year 2007, from x-axis which meet the trend line at point P. From P, a perpendicular on y-axis gives the required ternd estimate as 24.

Now, to find straight line trend, when number of years is even, consider the following example.

Example 8:

Fit a straight line trend to the following time –series, representing sales in lakhs of Rs. of a company, for the year 1998 to 2005. Plot the given data well as the trend line on a graph paper. Hence or otherwise estimate trend for the year 2006.

Year	1998	1999	2000	2001	2002	2003	2004	2005
Sales (Lakhs of Rs.)	31	33	30	34	38	40	45	49

Solution:

Here the number of years = 8, an even number, so we define

$x = \frac{year - 2001.5}{0.5}$, so that the values of x are -7, -3, -1, 1, 3, 5 and 7, to get $\sum x = 0$.

Prepare the following table to obtain the summations $\sum x^2$, $\sum y$, $\sum x y$.

Year	**Sales (in Lakhs of Rs.)**	**x**	**x^2**	**x y**	**Trend Values**
1998	31	-7	49	-217	28.33
1999	33	-5	25	-165	30.95
2000	30	-3	9	-90	33.57
2001	34	-1	1	-34	36.19
2002	38	1	1	38	38.81
2003	40	3	9	120	41.43
2004	45	5	25	225	44.05
2005	49	7	49	343	46.67
	300		168	220	

Here, n = 8, $\sum y = 300$, $\sum x^2 = 168$, $\sum x y = 220$

Now, a and b are calculated as follows:

$$a = \frac{\sum y}{n} = \frac{300}{8} = 37.5$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{220}{168} = 1.31$$

So, the equation of the straight line trend is $y = a + b x$

i.e. $y = 37.5 + 1.31 x$

To obtain the trend values, first calculate y for x = -7, for the year 1998

$y = 37.5 + 1.31 (-7)$

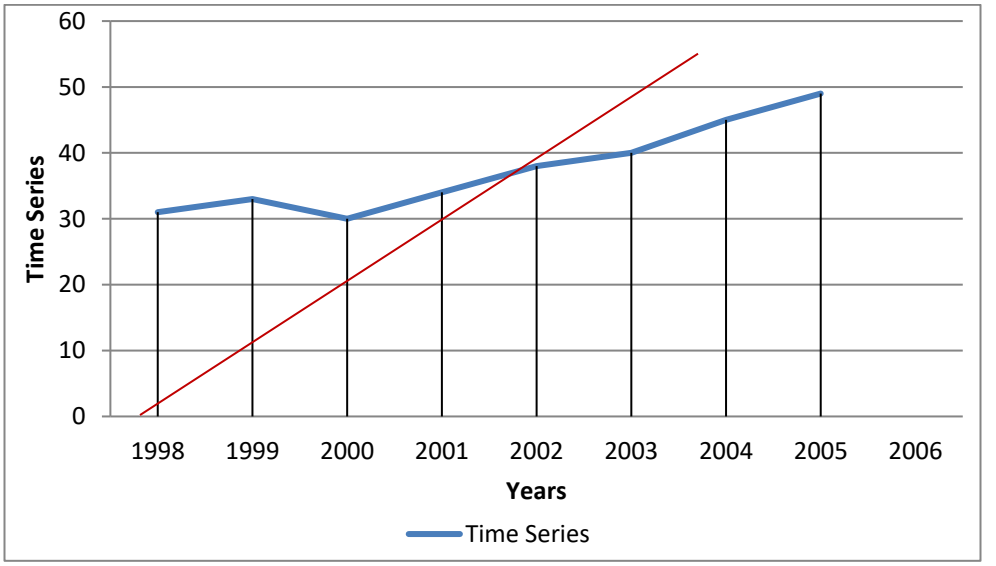
$= 37.5 – 9.17 = 28.33$

To find the successive trend values, go on addition $2b = 2 \times 1.31 = 2.62$, to the preceding values as in this case the different between x values is of 2 units.

So, the estimated values of trend for x = -5, -3, -1, 1, 3, 5, 7 and 7 are 30.95, 33. 57, 36.19, 38.81, 41.43, 44.05 and 46.67 respectively. Write down these values in the table.

Hence the estimated trend value for the year 2006 is 49.29 (in lakhs of Rs.).

Now, for the graph of trend line, note that only two trend values 30.95 and 46.67 w.r.t. years 1999 and 2005 are considered as point. The line joining these two points represents trend line.



To estimated the trend for the year 2006, drawn a perpendicular from x-axis at this point meeting the line in P. then from P, draw another perpendicular on y-axis which gives estimate of trend as 49.

Example 9:

Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the sales for the year 2007.

Year	1998	1999	2000	2001	2002	2003	2004	2005
Sales in'000Rs	120	124	126	130	128	132	138	137

Solution: let $y = a + bx$ be the straight line trend.

The number of years in the given time series is eight, which is an even number. The upper four years are assigned the values of x as 1, 2, 3, and 7. Note that ere the difference between the values of x is 2, but the sum is zero.

Now, the table of computation is completed as shown below:

Years	**Profit (y)**	**X**	**Xy**	**X^2**	**Trend Value: $Y_t = a + bx$**
1998	120	-7	-840	49	120.84
1999	124	-5	-620	25	123.28
2000	126	-3	-378	9	125.72
2001	130	-1	-130	1	128.16
2002	128	1	128	1	130.06
2003	132	3	396	9	133.04
2004	138	5	390	25	135.48
2005	137	7	359	49	137.92
Total	$\sum y = 1035$	$\sum x = 0$	$\sum xy = 205$	$\sum x^2 = 168$	

From the table : n = 8, $\sum xy = 205$, $\sum x^2 = 168$, $\sum y = 1035$

$$\therefore\ b = \frac{\sum xy}{\sum x^2} = \frac{205}{168} = 1.22 \quad \text{and} \quad a = \frac{\sum y}{n} = \frac{1035}{8} = 129.38$$

Thus, the straight line trend is y = 129.38 + 1.22x.

The trend values in the table for the respective years are calculated by substituting the corresponding value of x in the above trend line equation.

For the trend value for 1998: x = -7:

y_{1998} = 129.38 + 1.22 (-7) = 129.38- 8.54 = 120.84

Similarly, all the remaining trend values are calculated.

(A short-cut method in case of even number of years to find the remaining trend values once we calculate the first one, is to add twice the value of b to the first trend value to get the second trend value, then to the second trend value to get the third one and so on. This is because the difference in the values of x is 2. In this example we add 2 x 1.22 = 2.44)

Estimation:

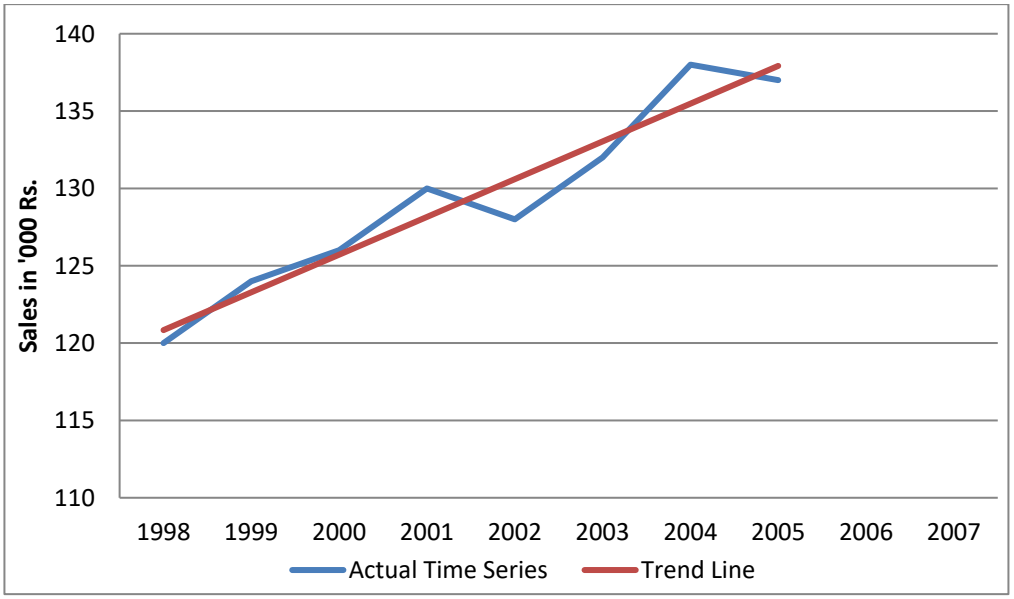
To estimate the profit for the years 2007 in the trend line equation, we substitute the prospective value of x, if the table was extended to 2007. i.e. we put x =11, the next value after x = 9 for the year 2006 and x = 7 for 2005.

$$y_{2007} = 129.38 + 1.22\ (11) = 142.8$$

There fore the estimated profit for the year 2007 is Rs. 1,42,800.

Now we draw the graph of actual time series by plotting the sales against the corresponding year, the period is taken on the X-axis and the sales on the Y-axis. The points are joined by straight lines. To draw the trend line it is enough to plot any two point (usually we take the first and the last trend value) and join it by straight line.

To estimate the trend value for the year 2007, we draw a line parallel to Y-axis from the period 2007 till it meet the trend line at a point say A. From this point we draw a line parallel to the X-axis till it meet the Y-axis at point say B. This point is our estimate value of sales for the year 2007. The graph and its estimate value (graphically) is shown below:



From the graph, the estimated value of the sales for the year 2007 is 142 i.e. Rs 1,42,000 (approximately)

Example 10:

Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the import for the year 1998.

	1991	1992	1993	1994	1995	1996
Import in'000Rs	40	44	48	50	46	52

Solution: Here again the period of years is 6 i.e. even. Proceeding similarly as in the above problem, the table of calculation and the estimation is as follows:

Years	**Import (y)**	**x**	**xy**	**x^2**	**Trend Value: $Y_t = a + bx$**
1991	40	-5	-200	25	41.82
1992	44	-3	-132	9	43.76
1993	48	-1	-48	1	45.7
1994	50	1	50	1	47.64
1995	46	3	138	9	49.58
1996	52	5	260	25	51.52
Total	$\sum y = 280$	$\sum x = 0$	$\sum xy = 68$	$\sum x^2 = 70$	

From the table : n = 6, $\sum xy = 68$, $\sum x^2 = 70$, $\sum y = 280$

Their four b = $\frac{\sum xy}{\sum x^2} = \frac{68}{70} = 0.97$ and $\frac{\sum y}{n} = \frac{280}{8} = 46.67$

Thus, the straight line trend is y = 46.67 + 0.97x.

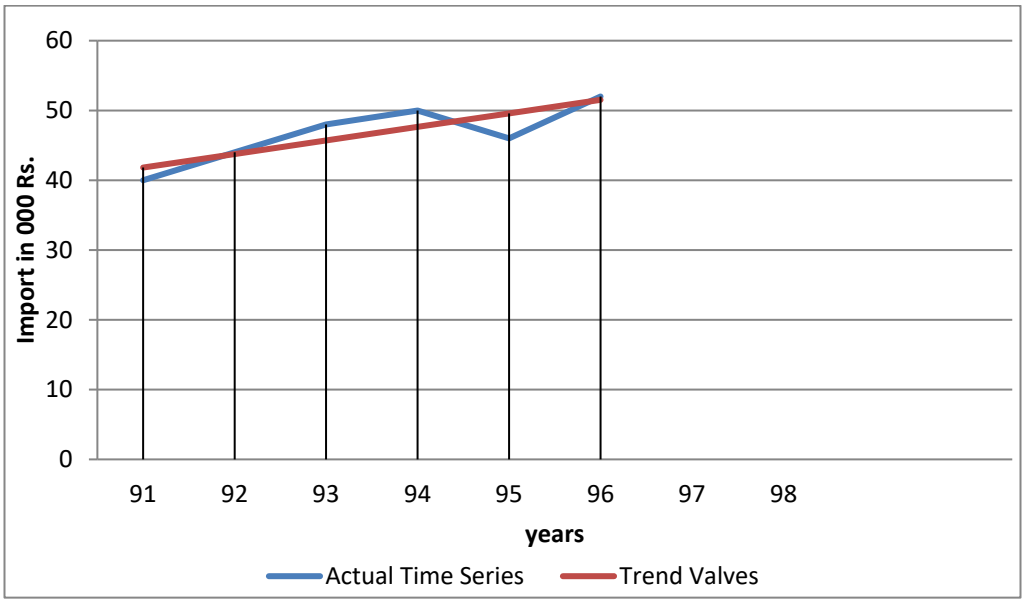
All the remaining trend values are calculated as described in the above problem.

Estimation:

To estimate the import for the year 1998, we put x = 9 in the tried line equation.There fore $y_{1997} = 46.67 + 0.97\ (9) = 55.4$

There fore the imports for the year 1997is Rs. 55,400.

The graph of the actual time series and the trend values along with the graphical estimation is an shown below:



From graph the estimated import are Rs. 55,000.

Example 11:

Fit a straight line trend to the following time –series, representing sales in lakhs of Rs. of a company, for the year 1998 to 2005. Plot the given data well as the trend line on a graph paper. Hence or otherwise estimate trend for the year 2006.

Year	1998	1999	2000	2001	2002	2003	2004	2005
Sales (Lakhs of Rs.)	31	33	30	34	38	40	45	49

Solution:

Here the number of years = 8, an even number, so we define $x = \frac{year - 2001.5}{0.5}$, so that the values of x are -7, -3, -1, 1, 3, 5 and 7, to get $\sum x = 0$.

Prepare the following table to obtain the summations $\sum x^2$, $\sum y$, $\sum x y$.

Year	**Sales (in Lakhs of Rs.)**	**x**	**x^2**	**x y**	**Trend Values**
1998	31	-7	49	-217	28.33
1999	33	-5	25	-165	30.95
2000	30	-3	9	-90	33.57
2001	34	-1	1	-34	36.19

Year	Sales (in Lakhs of Rs.)	x	x^2	x y	Trend Values
2002	38	1	1	38	38.81
2003	40	3	9	120	41.43
2004	45	5	25	225	44.05
2005	49	7	49	343	46.67
	300		168	220	

Here, n = 8, $\sum y = 300$, $\sum x^2 = 168$, $\sum x y = 220$

Now, a and b are calculated as follows:

$$a = \frac{\sum y}{n} = \frac{300}{8} = 37.5$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{220}{168} = 1.31$$

So, the equation of the straight line trend is y = a + b x

i.e. y = 37.5 + 1.31 x

To obtain the trend values, first calculate y for x = -7, for the year 1998

y = 37.5 + 1.31 (-7)

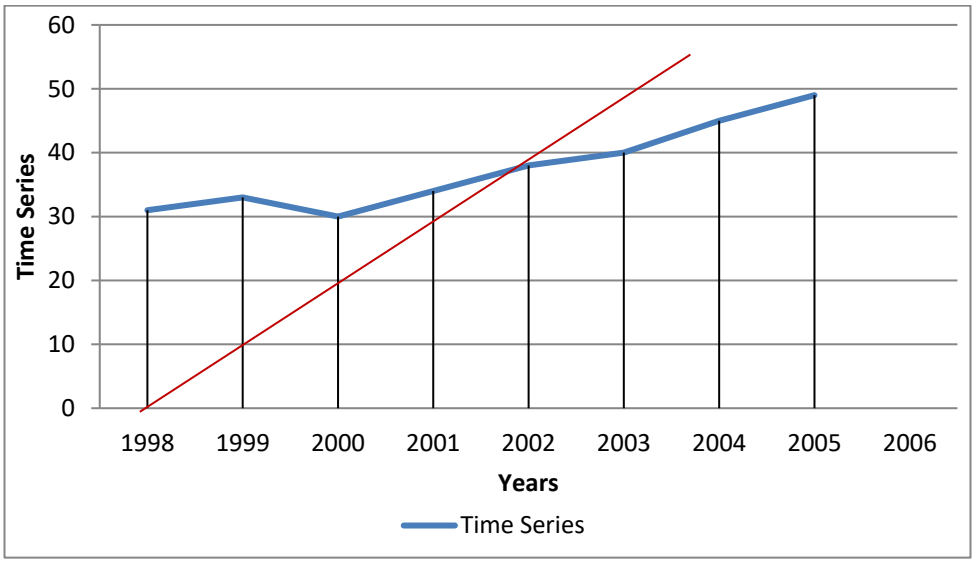
= 37.5 – 9.17 = 28.33

To find the successive trend values, go on addition 2b = 2 × 1.31 = 2.62, to the preceding values as in this case the different between x values is of 2 units.

So, the estimated values of trend for x = -5, -3, -1, 1, 3, 5, 7 and 7 are 30.95, 33. 57, 36.19, 38.81, 41.43, 44.05 and 46.67 respectively. Write down these values in the table.

Hence the estimated trend value for the year 2006 is 49.29 (in lakhs of Rs.).

Now, for the graph of trend line, note that only two trend values 30.95 and 46.67 w.r.t. years 1999 and 2005 are considered as point. The line joining these two points represents trend line.



To estimate the trend for the year 2006, drawn a perpendicular from x-axis at this point meeting the line in P. then from P, draw another perpendicular on y-axis which gives estimate of trend as 49.

MEASURESMENT OF OTHER COMPONENTS

We have studied four methods of estimation of Secular Trend. The following procedure is applied to separate the remaining components of the time series.

Using seasonal indices (s), the seasonal variations in a time series can be measured. By removing the trend and the seasonal factors, a combination of cyclical and irregular fluctuations is obtained.

If we assume, multiplicative model, represented by the equation

$$O = T \times S \times C \times I$$

Then, to depersonalize the data, the original time series (O) divided by the seasonal indices (S), which can be express as,

$$\frac{O}{S} = \frac{T \times S \times C \times I}{S} = T \times C \times I$$

If it is further divided by trend values (T), then we have

$$\frac{T \times C \times I}{T} = C \times I$$

Thus a combination of cyclical and irregular variation can be obtained. Irregular fluctuations, because of their nature, can not be eliminated completely, but these can be minimized by taking short term averages and then the estimate of cyclical variation can be obtained.

METHODS TO ESTIMATE SEASONAL FLUCTUATIONS

We have seen methods to separate the trend component of Time Series. Now, let us see, how to separate the seasonal component of it.

Methods of Seasonal Index

It is used to finds the effect of seasonal variations in a Time Series. The steps are as follows:

i. Find the totals for each season, as well as the grand total, say G.

ii. Find the arithmetic means of these total, and the grand total by dividing the values added.

iii. Find seasonal indices, representing the seasonal component for each season, using the formula

$$\text{Seasonal Index} = \frac{Average for Seasonal \times 100}{Grand Average}$$

$$\text{Where, Grand Average} = \frac{G}{Total No. of Values}$$

Example 12:

Find the seasonal component of the time series, using method of seasonal indices.

Seasonal / Years	I	II	IV	Grand
2003	33	37	32	31
2004	35	40	36	35
2005	34	38	34	32
2006	36	41	35	36
2007	34	39	35	32

Solution:

	I	II	III	IV	Grand
Total	172	195	172	166	705 (G)
Average	34.4 (172 / 5)	39	34.4	33.2	35.25(G/20)
Seasonal Index	$\frac{34.4 \times 100}{35.25}$ = 97.59	$\frac{39 \times 100}{35.25}$ = 110.64	$\frac{34.4 \times 100}{35.25}$ = 97.59	$\frac{33.2 \times 100}{35.25}$ = 94.18	

The time series can be deseasonalised by removing the effect of seasonal component from it. It is done using the formula.

$$\text{Deseasonalised Value} = \frac{OriginalValue \times 100}{SeasonalIndex}$$

BUSINESS FORECASTING:

In this chapter, few methods of analyzing the past data and predicting the future values are already discussed. Analysis of time series an important role in Business Forecasting. One of the aspects of it estimating future trend values. Now-a-days, any business or industry is governed by factors like supply of raw material, distribution network, availability of land, labour and capital and facilitates like regular supply of power, coal, water, etc. a business has to sustain intricate government regulations, status, everchanging tastes and fashions, the latest technology, cut throat competition by other manufacturers and many other.

While making a forecast, combined effect of above factors should be considered. Scientific method are used to analyse the past business condition. The study reveals the pattern followed by the business in the past. It also bring out the relationship and interdependence of different industries which helps in interpretation of changes in the right perspective. The analysis gives an idea about the components of the time series and their movement in the past. Various indices such as index of production, prices, bank deposits, money rates, foreign exchange position etc. can provide information about short and long term variations, the general trend, the ups downs in a business.

The study of the past data and the comparison of the estimated and actual values helps in pinpointing the areas of shortcoming which can be overcome. For successful business forescasting co-ordination of all departments such as production, sales, marketing is sine-qua-nin, which result in achieving ultimate corporate goals.

There are different theories of Business Forecasting such as

i. Time lag or Sequence Theory

ii. Action and Reaction Theory

iii. Cross Cut Analysis Theory

iv. Specific Historical Analogy Theory

Of these, Time lag or Sequence Theory is most important. It is based on the fact that there is a time lag between the effect of changes at different stages but there is

a sequence followed by these effect e.g. In 80's, the invention of silicon ships brought fourth and fifth generation computers in use. The computers were introduced in various fields such as front-line and back house banking, airlines and railways reservation, new communication technique, home appliances like washing machine etc. this, in turn, increase the demand for qualified personnel in electronic filed to manufacture, handle and maintain these sophisticated machine. It has result in mad rush for admission to various branches of electronics and computer engineering in the recent past.

By applying any one of the these forecasting theories, business forecasting can be made. It should be noted that while collecting the data for analysis, utmost care has to be taken so as to increase the reliability of estimates. The information should be collected by export investigators, over a long period of time. Otherwise, it may lead to wrong conclusions.

EXERCISE

1. What is a time series ? Describe the various components of a time series with suitable example.

2. What are seasonal variation ? Explain briefly with example.

3. Describe the secular trend component of a time series,

4. What are the method of determining trend in a time series?

5. Compare method of moving average and least squares of estimating trend component.

6. Find the trend values using the method of semi-averages for the following data expressing production in thousand unit of a company for 7 years.

7. Explain the method to calculate 3 yearly and 4 yearly moving averages.

8. What are the merits and demerits of the method of moving average?

9. Explain the simple average method to find the seasonal indices of a time series

10. Calculate trend by considering three yearly moving average for the following time series of price indices for the years 2000-2007. Also plot on the graph the trend values.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Price Index	111	115	116	118	119	120	122	124

11. Determine the trend for the following data using 3 yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997
Sales in'000Rs	24	28	30	33	34	36	35	40	44

12. Determine the trend for the following data using 3 yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1977	1978	1979	1980	1981	1982	1983	1984
Sales in'000Rs	46	54	52	56	58	62	59	63

13. Determine the trend for the following data using 3 yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1979	1980	1981	1982	1983	1984	1985	1986
Profit in lakhs of Rs	98	100	97	101	107	110	102	105

14. Determine the trend for the following data using 5 yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1980	1982	1984	1986	1988	1990	1992	1994	1996	1998	2000
Values	34	37	35	38	37	40	43	42	48	50	52

15. Determine the trend for the following data giving the production of steel in million tons, using 5 yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
Production	28	30.5	32	36.8	38	36	39.4	40.6	42	45	43.5

17. Find five-yearly moving average for the following data which represents production in thousand unit of a small scale industry. Plot the given data as well as the moving average on a graph paper.

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Production	110	104	78	105	109	120	115	110	115	122	130

Ans. The trend values are 101.2, 103.2, 105.4, 111.8, 113.8, 116.4 and 118.4 for the years 1982 to 1988.

18. Find the trend component of the following time series of production in thousand kilogram during 1971-1980. Plot the moving average and the original time on a graph paper.

Year	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Production	12	15	18	17	16	20	23	22	24	25

Ans. The trend values are 16, 17.125, 18.375, 19.625, 21.25, 22.875 for the years 1973 to 1978.

19. Fit a straight line trend to the following data representing import in million Rs. of a certain company. Also find an estimate for the year 2008.

Year	2000	2001	2002	2003	2004	2005	2006
Import	48	50	58	52	45	41	49

Ans. The straight line trend is y = 49-x. the trend values are 52, 51, 50, 49, 48, 47 and 46 respectively and the estimate trend for the year 2006 is 44 million Rs.

20. The production of a certain brand of television sets in thousand unit is given below. Fit a straight line trend to the data. Plot the given data and the trend line on graph find an estimate for the year 2004.

Year	1997	1998	1999	2000	2001	2002	2003
production	865	882	910	925	965	1000	1080

Ans. The straight line trend is y = 947.71 + 33.43 x. the trend values are 846.42, 879.85, 913.28, 946.71, 1013.57 and 1047. The estimate for the year 2004 is 1080.43 thermal million.

21. The straight line trend by the method of least squares for the following data which represents the expenditure in lakhs od Rs. on advertisement of a certain company. Also find an estimate for the year 2005. Plot the given data and the trend line on a graph paper.

Year	1997	1998	1999	2000	2001	2002	2003	2004
Expenditure	21	24	32	40	38	49	57	60

Ans. The trend is y = 40.13 +2.9x. the trend values are 19.83, 25.62, 31.43, 37.23, 43.03, 48.83, 54.63 and 60.43, 2005 is 66.23.

22. Use the method of least squares to find straight line trend for the following time series of production in thousand units 1981 – 1988. Also estimate trend for the year 2003.

Year	1995	1996	1997	1998	1999	2000	2001	2002
Production	80	90	92	83	94	99	92	102

Ans. The straight line trend is y = 91.5 + 1.167 x. the trend values are 83.331, 85.665, 87.999, 90.333, 92.667, 95.001, 97.335 and 99.669. the estimate of trend, for the year 2003 is 102.003

23. Calculate seasonal indices for the following data:

Year	I	II	III	IV
2003	55	53	57	51
2004	56	55	60	53
2005	57	56	61	54

Ans. 100.59, 98.2, 106.57, 94.61

24. Determine the trend for the following data giving the production of wheat in thousand tons from the years 1980 to 1990, using the 5-yearly moving averages. Plot the graph of actual time series and the trend values.

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Production	13.5	14.7	17	16.2	18.1	20.4	22	21.2	24	25	26.6

25. Determine the trend for the following data giving the income (in million dollars) from the export of a product from the year 1988 to 1999. Use the 4-yearly moving average method and plot the graph of actual time series and trend values.

Year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Income	340	360	385	470	430	444	452	473	490	534	541	576

26. Using the 4-yearly moving average method find the trend for the following data.

Year	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977
Value	102	100	103	105	104	109	112	115	113	119	117

27. Determine the trend for the following data giving the sales (in ’00 Rs.) of a product per week for 20 weeks. Use appropriate moving average method.

Week	1	2	3	4	5	6	7	8	9	10
Sales	22	26	28	25	30	35	39	36	30	32
Week	11	12	13	14	15	16	17	18	19	20
Sales	29	34	36	35	35	39	43	48	52	49

28. An online marketing company works 5-days a week. The day-to-day total sales (in '000 Rs) of their product for 4 weeks are given below. Using a proper moving average method find the trend values.

Days	1	2	3	4	5	6	7	8	9	10
Sales	12	16	20	17	18	20	26	25	27	30
Days	11	12	13	14	15	16	17	18	19	20
Sales	35	32	32	38	36	35	34	38	40	41

29. Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the sales for the years 2000.

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Sales in '000 Rs	45	47	49	48	54	58	53	59	62	60	64

30. Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the sales for the years 2001.

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Profit in '000 Rs	76	79	82	84	81	84	89	92	88	90

31. Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the sales for the years 2007.

Year	1998	1999	2001	2002	2003	2004	2005	2006
Profit in '000 Rs	116	124	143	135	138	146	142	152

32. Fit a straight line trend to the following data giving the number of casualties (in hundred) of motorcyclists without helmet. Estimate the number for the year 1999.

Year	1992	1993	1994	1995	1996	1997	1998
No of casualties	12	14.2	15.2	16	18.8	19.6	22.1

33. Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the import for the years 2002

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Import in'000 Rs.	55	52	50	53	54	56	58	60	57	59

34. Fit a straight line trend to the following data giving the price of crude oil per barrel in USD. Draw the graph of the actual time series and the trend line. Estimate the sales for the year 2003.

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Price per barrel	98	102	104.5	108	105	109	112	118	115	120

35. Apply the method of least squares to find the number of student attending the library in the month of May of the academic year 2005 – 2006 from the following data.

Month	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr
Students	105	120	160	225	180	115	124	138	176	230	180

36. Assuming that the trend is absent, find the seasonal indices for the following data and also find the deseasonalized values.

Quarters	**I**	**II**	**III**	**IV**
1977	10	12	14	16
1978	12	15	18	22
1979	16	18	20	24
1980	24	26	28	34

37. Calculate seasonal indices for the following data:

Year	**I**	**II**	**III**	**IV**
2003	55	53	57	51
2004	56	55	60	53
2005	57	56	61	54

Ans. 100.59, 98.2, 106.57, 94.61

UNIT 4

10

CORRELATION AND REGRESSION

Unit Structure

10.1 Objectives

- To understand the relationship between two relevant characteristics of a statistical unit.
- Learn to obtain the numerical measure of the relationship between two variables.
- Use the mathematical relationship between two variables in order to estimate value of one variable from the other.
- Use the mathematical relationship to obtain the statistical constants line means and S.D.'s

10.1 Introduction

In the statistical analysis we come across the study of two or more relevant characteristics together in terms of their interrelations or interdependence. e.g. Interrelationship among production, sales and profits of a company. Inter relationship among rainfall, fertilizers, yield and profits to the farmers.

Relationship between price and demand of a commodity When we collect the information (data) on two of such characteristics it is called bivariate data. It is generally denoted by (X,Y) where X and Y are the variables representing the values on the characteristics.

Following are some examples of bivariate data.

a) Income and Expenditure of workers.

b) Marks of students in the two subjects of Maths and Accounts.

c) Height of Husband and Wife in a couple.

d) Sales and profits of a company.

Between these variables we can note that there exist some sort of interrelationship or cause and effect relationship. i.e. change in the value of one variable brings out the change in the value of other variable also. Such relationship is called as correlation.

Therefore, correlation analysis gives the idea about the nature and extent of relationship between two variables in the bivariate data.

10.2 Types of Correlation

There are two types of correlation.

a) Positive correlation. and
b) Negative correlation.

10.2.1 Positive correlation: When the relationship between the variables *X* and Y is such that increase or decrease in X brings out the increase or decrease in *Y* also, i.e. there is direct relation between X and Y, the correlation is said to be positive. In particular when the 'change in X equals to change in *Y*' the correlation is perfect and positive. e.g. Sales and Profits have positive correlation.

10.2.2 Negative correlation: When the relationship between the variables X and Y is such that increase or decrease in X brings out the decrease or increase in Y, i.e. there is an inverse relation between X and Y, the correlation is said to be negative. In particular when the 'change in X equals to change in Y' but in opposite direction the correlation is perfect and negative. e.g. Price and Demand have negative correlation.

10.3 Measurement o f Correlation

The extent of correlation can be measured by any of the following methods:

- Scatter diagrams
- Karl Pearson's co-efficient of correlation
- Spearman's Rank correlation

10.3.1 Scatter Diagram: The Scatter diagram is a chart prepared by plotting the values of X and Y as the points (X,Y) on the graph. The pattern of the points is used to explain the nature of correlation as follows.

The following figures and the explanations would make it clearer.

(*i*) *Perfect Positive Correlation*:

If the graph of the values of the variables is a straight line with positive slope as shown in Figure **10.1**,

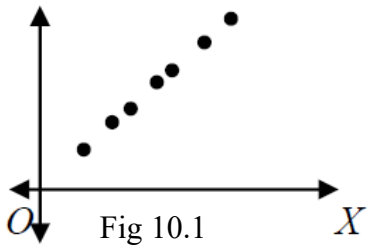


Fig 10.1

we say there is a *perfect positive correlation* between *X* and *Y*. Here $r = 1$.

(*ii*) *Imperfect Positive Correlation*:

If the graph of the values of *X* and *Y* show a band of points from lower left corner to upper right corner as shown in Figure **10.2**,

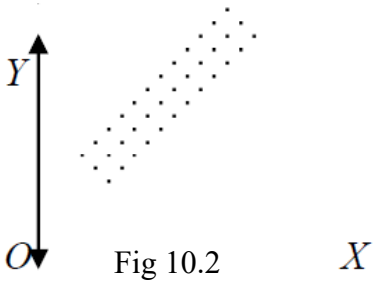


Fig 10.2

we say that there is an *imperfect positive correlation*. Here $0 < r < 1$.

(*iii*) *Perfect Negative Correlation*:

If the graph of the values of the variables is a straight line with negative slope as shown in Figure **10.3**,

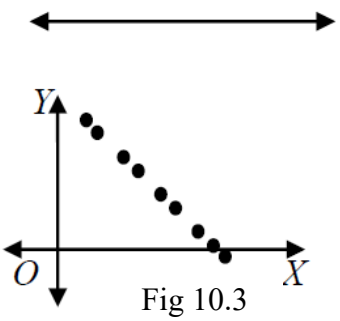


Fig 10.3

we say there is a *perfect negative correlation* between *X* and *Y*. Here $r = -1$.

(*iv*) *Imperfect Negative Correlation*: If the graph of the values of X and Y show a band of points from upper left corner to the lower right corner as shown in Figure **10.4**,

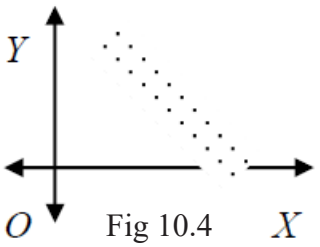


Fig 10.4

then we say that there is an *imperfect negative correlation*. Here $-1 < r < 0$

(*v*) *Zero Correlation*:

If the graph of the values of X and Y do not show any of the above trend then we say that there is a *zero correlation* between X and Y. The graph of such type can be a straight line perpendicular to the axis, as shown in Figure **10.5** and **10.6,** or may be completely scattered as shown in Figure **10.7**. Here $r = 0$.

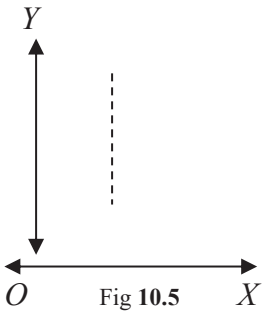


Fig **10.5**

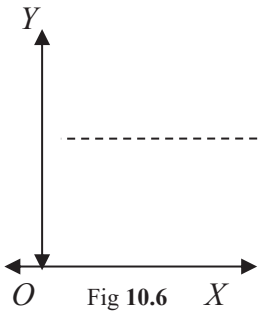
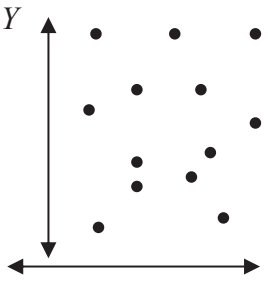


Fig **10.6**



The Figure **10.5** show that the increase in the values of Y has no effect on the value of X, it remains the same, hence zero correlation. The Figure **4.6** show that the increase in the values of X has no effect on the value of Y, it remains the same, hence zero correlation. The Figure **10.7** show that the points are completely scattered on the graph and show no particular trend, hence there is no correlation or zero correlation between X and Y.

10.3.2 Karl Pearson's co-efficient of correlation.

This co-efficient provides the numerical measure of the correlation between the variables X and Y. It is suggested by Prof. Karl Pearson and calculated by the formula

$$r = \frac{Cov(x, y)}{\sigma_x . \sigma_y}$$

Where, Cov(x,y) : Covariance between *x* & *y*

σ_x.: Standard deviation of *x* & σ_y: Standard deviation of *y*

Also, $\text{Cov}(x,y) = \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y}) = \frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}$

$$\text{S.D.}(x) = \sigma x = \sqrt{\frac{1}{n}\Sigma(x-\bar{X})^2} = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{X}^2} \quad \text{and}$$

$$\text{S.D.}(y) = \sigma y = \sqrt{\frac{1}{n}\Sigma(y-\bar{Y})^2} \qquad \sqrt{\frac{1}{n}\Sigma y^2 - \bar{Y}^2}$$

Remark : We can also calculate this co-efficient by using the formula given by

$$r = \frac{\frac{1}{n}\Sigma(x-\bar{X})(y-\bar{Y})}{\sqrt{\frac{1}{n}\Sigma(x-\bar{X})^2}\sqrt{\frac{1}{n}\Sigma(y-\bar{Y})^2}} = \frac{\frac{\Sigma xy}{n} - \bar{X}\bar{Y}}{\sqrt{\left(\frac{\Sigma x^2}{n} - \bar{X}^2\right)\left(\frac{\Sigma y^2}{n} - \bar{Y}^2\right)}}$$

The Pearson's Correlation co-efficient is also called as the 'product moment correlation co-efficient'

Properties of correlation co-efficient 'r'

The value of 'r' can be positive (+) or negative(-)

The value of 'r' always lies between –1 & +1, i.e. –1< r<+1]

Significance of 'r' equals to –1, +1 & 0

When 'r'= +1; the correlation is perfect and positive.

'r'= -1; the correlation is perfect and negative.

and when there is no correlation 'r'= 0

SOLVED EXAMPLES :

Example.1: Calculate the Karl Pearson's correlation coefficient from the following.

X:	12	10	20	13	15
Y:	7	14	6	12	11

Solution: Table of calculation,

X	Y	XxY	X^2	Y^2
12	7	84	144	49
10	14	140	100	196
20	6	120	400	36
13	12	156	169	144
15	11	165	225	121
$\underline{\Sigma x}$ = 70	$\underline{\Sigma y}$ = 50	$\underline{\Sigma xy}$= 665	$\underline{\Sigma x^2}$ =1038	$\underline{\Sigma y^2}$ =546

And n= 5

The Pearson's correlation coefficient r is given by,

$$r = \frac{Cov(x,y)}{\sigma_x . \sigma_y}$$

Where,

$$\bar{x} = \frac{\Sigma x}{n} = \frac{70}{5} = 14 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{50}{5} = 10$$

$$\text{Cov(x,y)} = \frac{\Sigma xy}{n} - \bar{X}\bar{Y} \qquad \sigma_x = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{X}^2} \qquad \sigma_y = \sqrt{\frac{1}{n}\Sigma y^2 - \bar{Y}^2}$$

$$= \frac{665}{5} - 14x10 \qquad = \sqrt{\frac{1038}{5} - 14^2} \qquad = \sqrt{\frac{546}{5} - 10^2}$$

$$= 133 - 140 \qquad = \sqrt{11.6} = 3.40 \quad \sqrt{9.2} = 3.03 = -07$$

∴ Cov(x,y) = -7 σ_x = 3.40 and σ_y = 3.03

Substituting the values in the formula of **r** we get

$$r = \frac{-7}{3.40x3.03} = -0.68$$

.

∴ r = - 0.68

Example 2: Let us calculate co-efficient of correlation between Marks of students in the Subjects of Maths & Accounts. in a certain test conducted.

Table of calculation:

Marks InMaths X	Marks In Accounts Y	XY	X^2	Y^2
28	30	840	784	900
25	40	1000	625	1600
32	50	1600	1024	2500
16	18	288	256	324
20	25	500	400	625
15	12	180	225	144
19	11	209	361	121
17	21	357	289	441
40	45	1800	1600	2025
30	35	1050	900	1225
Σx = 242	Σy=287	Σxy= 7824	Σx^2 = 6464	Σy^2 9905

n=10

Now Pearson's co-efficient of correlation is given by the fomula,

$$r = \frac{Cov(x, y)}{\sigma_x . \sigma_y}$$

Where,

$$\bar{x} = \frac{\Sigma x}{n} = \frac{242}{10} = 24.2 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{287}{10} = 28.7$$

$$\text{Cov(x,y)} = \frac{\Sigma xy}{n} - \bar{X}\bar{Y}$$

$$= \frac{7824}{10} - 24.2 \times 28.7$$

$$= 782.4 - 694.54$$

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{X}^2}$$

$$\sigma_x = \sqrt{\frac{6464}{10} - 24.2^2}$$

$$= \sqrt{60.76}$$

$$\sigma_y = \sqrt{\frac{1}{n}\Sigma y^2 - \bar{Y}^2}$$

$$\sigma_y = \sqrt{\frac{9905}{10} - 28.7^2}$$

$$= \sqrt{166.81}$$

Cov(x,y) = 87.86, σ_x = 7.79 and σ_y = 12.91

∴ Cov(x,y) = 87.86 σ_x = 7.79 and σ_y = 12.91

Substituting the values in the formula of **r** we get

$$r = \frac{87.86}{7.79 x 12.91} = \mathbf{0.87}$$

∴ r = 0.87

4.4 RANK CORRELATION

In many practical situations, we do not have the scores on the characteristics, but the ranks (preference order) decided by two or more observers. Suppose, a singing competition of 10 participants is judged by two judges A and B who rank or assign scores to the participants on the basis of their performance. Then it is quite possible that the ranks or scores assigned may not be equal for all the participants. Now the difference in the ranks or scores assigned indicates that there is a difference of openion between the judges on deciding the ranks. The rank correlation studies the association in this ranking of the observations by two or more observers. The measure of the extent of association in rank allocation by the two judges is calculated by the co-efficient of Rank correlation 'R'. This co-efficient was developed by the British psychologist Edward Spearman in 1904.

Mathematically, Spearman's rank correlation co-efficient is defined as,

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Where d= rank difference and n= no of pairs.

Remarks: We can note that, the value of 'R' always lies between –1 and +1
The positive value of 'R' indicates the positive correlation (association) in the rank allocation. Whereas, the negative value of 'R' indicates the negative correlation (association) in the rank allocation.

SOLVED EXAMPLES:

Example 3

a) <u>When ranks are given:-</u>

Data given below read the ranks assigned by two judeges to 8 participants. Calculate the co-efficient of Rank correlation.

Participant No.	Ranks by Judge		Rank diff Square d^2
	A	B	
1	5	4	$(5-4)^2 = 1$
2	6	8	4
3	7	1	36
4	1	7	36
5	8	5	9
6	2	6	16
7	3	2	1
8	4	3	1
N = 8	Total		$104 = \Sigma d^2$

Spearman's rank correlation co-efficient is given by

$$R=1-\frac{6\Sigma d^2}{n(n^2-1)}$$

Substituting the values from the table we get,

$$R= 1-\frac{6x\ 104}{8(8^2-1)} = -0.23$$

The value of correlation co-efficient is - 0.23. This indicates that there is negative association in rank allocation by the two judges A and B

b) <u>When scores are given:-</u>

Example 4

The data given below are the marks given by two Examiners to a set of 10 students in a aptitude test. Calculate the Spearman's Rank correlation co-efficient, 'R'

Now the Spearman's rank correlation co-efficient is given by

$$R= 1-\frac{6\Sigma d^2}{n(n^2-1)}$$

Substituting the values from the table we get,

$$R= 1-\frac{6x5}{10(10^2-1)}$$

$$= 1-0.03$$

$$= 0.97$$

The value of correlation co-efficient is + 0.97. This indicates that there is positive association in assessment of two examiners, A and B.

c) <u>Case of repeated values:-</u>

It is quite possible that the two participants may be assigned the same score by the judges. In such cases Rank allocation and calculation of rank correlation can be explained as follows.

Example: The data given below scores assigned by two judges for 10 participants in the singing competition. Calculate the Spearman's Rank correlation co-efficient.

Participant No.	Score assigned By Judges		Ranks		Rank difference square
	A	B	R_A	R_B	D^2
1	28	35	**9 (8.5)**	6	$(8.5-6)^2$ =6.25
2	40	26	3	**10(9.5)**	42.25
3	35	42	**5 (4.5)**	3	2.25
4	25	26	10	**9 (9.5)**	0.25
5	28	33	**8 (8.5)**	7	2.25
6	35	45	**4 (4.5)**	2	6.25
7	50	32	1	8	49
8	48	51	2	1	1
9	32	39	6	4	4
10	30	36	7	5	4
N = 10				Total	Σd^2 =117.5

Student No.	Marka By Examiner		Ranks		Rank difference square
	A	B	R_A	R_B	D^2
1	85	80	2	2	0
2	56	60	8	7	1
3	45	50	10	10	0
4	65	62	6	6	0
5	96	90	1	1	0
6	52	55	9	8	1
7	80	75	3	4	1
8	75	68	5	5	0
9	78	77	4	3	1
10	60	53	7	9	1
N = 10				Total	5 = Σd^2

Explanation:- In the column of A and B there is repeatation of scores so while assigning the ranks we first assign the ranks by treating them as different values and then for rereated scores we assign the average rank. e.g In col A the score 35 appears 2 times at number 4 and 5 in the order of ranking so we calculate the average rank as (4+5)/2 = 4.5.

Hence the ranks assigned are 4.5 each. The other repeated scores can be ranked in the same manner.

Note: In this example we can note that the ranks are in fraction e.g. 4.5, which is logically incorrect or meaningless. Therefore in the calculation of 'R' we add a correction factor (C.F.) to Σd^2 calculated as follows.

Table of correction factor (C.F.)

Value Repeated	Frequency M	$m(m^2-1)$
35	2	$2x(2^2-1)=6$
28	2	6
26	2	6
	Total	$\Sigma m(m^2-1)=18$

Now $C.F. = \frac{\Sigma(m^3 - m)}{12} = 18/12 = 1.5$

$\therefore \Sigma d^2 = 117.5+1.5= 119$

We use this value in the calculation of 'R'

Now the Spearman's rank correlation co-efficient is given by

$$R= 1-\frac{6\Sigma d^2}{n(n^2-1)}$$

Substituting the values we get, $R= 1- \frac{6x\ 119}{10(10^2-1)} = 1-0.72 = 0.28$

EXERCISE I

1. What is mean by correlation? Explain the types of correlation with suitable examples.

2. What is a scatter diagram? Draw different scattered diagrams to explain the correlation between two variables x and y.

3. State the significance of 'r' = +1, –1 and 0.

4. Calculate the coefficient of correlation r from the following data.

X: 18 12 16 14 10 15 17 13
Y: 9 13 20 15 11 24 26 22

5. The following table gives the price and demand of a certain commodity over the period of 8 months. Calculate the Pearson's coefficient of correlation.

Price:	15	12	23	25	18	17	11	19
Demand	45	30	60	65	48	45	28	50

6. Following results are obtained on a certain bivariate data.

(i) $n = 10$ $\Sigma x = 75$ $\Sigma y = 70$ $\Sigma x^2 = 480$
$\Sigma y^2 = 600$ $\Sigma xy = 540$

(ii) $n = 15$ $\Sigma x = 60$ $\Sigma y = 85$ $\Sigma x^2 = 520$
$\Sigma y^2 = 1200$ $\Sigma xy = -340$

Calculate the Pearson's correlation coefficient in each case.

7. Following data are available on a certain bi-variate data :

(i) $\Sigma(x-\bar{x})(y-\bar{y}) = 120$, $\Sigma(x-\bar{x})^2 = 150$ $\Sigma(y-\bar{y})^2 = 145$

(ii) $\Sigma(x-\bar{x})(y-\bar{y}) = -122$, $\Sigma(x-\bar{x})^2 = 136$ $\Sigma(y-\bar{y})^2 = 148$

Find the correlation coefficient.

8. Calculate the Pearson's coefficient of correlation from the given information on a bivariate series:

No of pairs: 25

Sum of x values:300

Sum of y values:375

Sum of squares of x values: 9000

Sum of squares of y values:6500

Sum of the product of x and y values:4000.

9. The ranks assigned to 8 participants by two judges are as followes.

Calculate the Spearman's Rank correlation coefficient 'R'.

Participant No:	1	2	3	4	5	6	7	8
Ranks by JudgeI:	5	3	4	6	1	8	7	2
JudgeII :	6	8	3	7	1	5	4	2

10. Calculate the coefficient of rank correlation from the data given below.

X:	40	33	60	59	50	55	48
Y:	70	60	85	75	72	82	69

11. Marks given by two Judges to a group of 10 participants are as follows. Calculate the coefficient of rank correlation.

Marks by

Judge A: 52 53 42 60 45 41 37 38 25 27

Judge B: 65 68 43 38 77 48 35 30 25 50.

12. An examination of 8 applicants for a clerical post was by a bank. The mar obtained by the applicants in the subjects of Mathematics and Accountancy we as follows. Calculate the rank correlation coefficient.

Applicant:	A	B	C	D	E	F	G	H
Marks in Maths:	15	20	28	12	40	60	20	80
Marks in Accounts:	40	30	50	30	20	10	25	60

10.5 Regression Analysis

As the correlation analysis studies the nature and extent of interrelationship between the two variables X and Y, regression analysis helps us to estimate or approximate the value of one variable when we know the value of other variable. Therefore we can define the 'Regression' as the estimation (prediction) of one variable from the other variable when they are correlated to each other. e.g. We can estimate the Demand of the commodity if we know it's Price.

Why are there two regressions?

When the variables X and Y are correlated there are two possibilities,

(i) Variable X depends on variable y. in this case we can find the value of x if know the value of y. This is called regression of x on γ.

(ii) Variable γ depends on variable X. we can find the value of y if know the value of X. This is called regression of y on x. Hence there are two regressions,

(a) Regression of X on Y; (b) Regression of *X* on Y.

10.5.1 Formulas on Regression equation,

Regression of X **on** Y	**Regression of** X **on** Y
Assumption: X depends on Y	Y depends on X
The regression equation is	The regression equation is
$(x-\bar{x}) = b_{xy}(y-\bar{y})$	$(y-\bar{y}) = b_{yx}(x-\bar{x})$
b_{xy}= Regression co-efficient of X on Y $= \frac{Cov(x,y)}{V(y)}$	b_{yx}= Regression co-efficient of Y on X $= \frac{Cov(x,y)}{V(x)}$

Where,

$$\text{Cov}(x,y) = \frac{1}{n}\underline{\Sigma}(x-\bar{x})(y-\bar{y}) = \frac{1}{n}\underline{\Sigma}xy-\bar{x}\,\bar{y}$$

$$V(x) = \frac{1}{n}\Sigma(x-\bar{x})^2 \quad \text{and} \quad V(y) = \frac{1}{n}\Sigma(y-\bar{y})^2$$

$$V(x) = \frac{1}{n}\Sigma x^2 - \bar{x}^2 \quad \text{and} \quad V(y) = \frac{1}{n}\Sigma y^2 - \bar{y}^2$$

Use: To find X Use: To find γ

SOLVED EXAMPLES

Example 1:

Obtain the two regression equation and hence find the value of x when y=25
Data:-

X	Y	X^2	Y^2	XxY
8	15	64	225	120
10	20	100	400	200
12	30	144	900	360
15	40	225	1600	600
20	45	400	2025	900
$\underline{\Sigma x}$= 65	$\underline{\Sigma y}$ =150	$\underline{\Sigma x^2}$= 933	$\underline{\Sigma y^2}$= 5150	$\underline{\Sigma xy}$= 2180

And n= 5

Now the two regression equations are,

$(x-\bar{x}) = b_{xy}(y-\bar{y})$-------x on y (i)

$(y-\bar{y}) = b_{yx}(x-\bar{x})$ ------- y on x (ii)

Where,

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{65}{5} = 13 \quad \text{and} \quad \bar{y} = \frac{1}{n}\Sigma y = \frac{150}{5} = 30$$

Also,

$$\text{Cov}(x,y,) = \frac{1}{n}\Sigma xy-\bar{x}\,\bar{y} = \frac{2180}{5} - 13\text{x}30 = 436-390$$

$$V(x) = \frac{1}{n}\Sigma x^2 - \bar{x}^2 = \frac{933}{5} - 13^2 = 186.6 - 169$$

$$V(y) = \frac{1}{n}\Sigma y^2 - \bar{y}^2 = \frac{5150}{5} - 30^2 = 1030 - 900$$

∴ Cov(x,y) = 46 V(x) = 17.6 V(y) =130

Now we find,

Regression co-efficient of X on Y	Regression co-efficient of X on Y
$b_{xy} = \frac{Cov(x,y)}{V(y)}$	$b_{yx} = \frac{Cov(x,y)}{V(x)}$
$= \frac{46}{130}$	$= \frac{46}{17.6}$
$\therefore\ b_{xy} = 0.35$ and	$b_{yx} = 2.61$

Now substituting the values of $\bar{x}$, $\bar{y}$, b_{xy} and b_{yx} in the regression equations we get,

(x-13) = 0.35(y-30) -------x on y (i)
(y-30) =2.61(x-13) ------- y on x (ii)
as the two regression equations.

Now to estimate x when y =25, we use the regression equation of x on y

$\therefore$ (x-13) = 0.35(25-30)
$\therefore$ **x = 13 -1.75 = 11.25**

Remark:

From the above example we can note some points about Regression coefficients.

- Both the regression coefficients carry the same sign (+ or -)
- Both the regression coefficients can not be greater than 1 in number (e.g. -1.25 and -1.32) is not possible.
- Product of both the regression coefficients b_{xy} and b_{yx} must be < 1
  i.e. $b_{xy} \times b_{yx} < 1$ Here 0.35x2.61 = 0.91< 1 ***(Check this always)***

Example 2:

Obtain the two regression equations and hence find the value of y when x=10
Data:-

X	Y	XxY	X^2	Y^2
12	25	300	144	625
20	18	360	400	324
8	17	136	64	289
14	13	182	196	169
16	15	240	256	225
Σx= 70	Σy=88	Σxy=1218	Σx^2=1060	Σy^2=1632

And n= 5

Now the two regression equations are,

$(x-\bar{x}) = b_{xy}(y-\bar{y})$-------x on y (i)

$(y-\bar{y}) = b_{yx}(x-\bar{x})$ ------- y on x (ii)

Where,

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{70}{5} = 14 \qquad \text{and} \qquad \bar{y} = \frac{1}{n}\Sigma y = \frac{88}{5} = 17.6$$

Also,

$$\text{Cov}(x,y,) = \frac{1}{n}\Sigma xy - \bar{x}\,\bar{y} = \frac{1218}{5} - 14\text{x}17.6 = 243.6 - 246.4$$

$\therefore$ **Cov(x,y)** = -2.8

$$V(x) = \frac{1}{n}\Sigma x^2 - \bar{x}^2 = \frac{1060}{5} - 14^2 = 212 - 196$$

V(x) = 16

$$V(y) = \frac{1}{n}\Sigma y^2 - \bar{y}^2 = \frac{1632}{5} - 17.6^2 = 326.4 - 309.76$$

V(y) = 16.64

Now we find,

Regression co-efficient of X on Y

$$b_{xy} = \frac{Cov(x,y)}{V(y)} = \frac{2.8}{16.64}$$

$\therefore$ b_{xy} = - 0.168

Regression co-efficient of X on Y

$$b_{yx} = \frac{Cov(x,y)}{V(x)} = -\frac{2.8}{16.64}$$

b_{yx} = 0.175

Now substituting the values of $\bar{x}$, $\bar{y}$, b_{xy} and b_{yx} in the regression equations we get,

$(x-14) = -0.168(y-17.6)$ -------x on y (i)

$(y-17.6) = -0.175(x-14)$ ------- y on x (ii)

as the two regression equations.

Now to estimate y when x =10, we use the regression equation of y on x

$\therefore$ $(y-17.6) = -0.175(10-14)$

$\therefore$ **y =17.6 +0.7 = 24.3**

Example 3:

The following data give the experience of machine operators and their performance rating given by the number of good parts turned out per 100 pieces.

Operator:	1	2	3	4	5	6	7	8
Experience: (in years)	16	12	18	4	3	10	5	12
Performance: Rating	87	88	89	68	78	80	75	83

Obtain the two regression equations and estimate the permance rating of an operator who has put 15 years in service.

Solution: We define the variables,
X: Experience y: Performance rating

Table of calculations:

X	Y	Xy	x^2	Y^2
16	87	1392	256	7569
12	88	1056	144	7744
18	89	1602	324	7921
4	68	272	16	4624
3	78	234	9	6084
10	80	800	100	6400
5	75	375	25	5625
12	83	996	144	6889
$\Sigma x = 80$	$\Sigma y = 648$	$\Sigma xy = 6727$	$\Sigma x^2 = 1018$	$\Sigma y^2 = 52856$

Now the two regression equations are,

$$(x - \bar{x}) = b_{xy}(y - \bar{y}) \quad \text{-------x on y (i)}$$

$$(y - \bar{y}) = b_{yx}(x - \bar{x}) \quad \text{------- y on x (ii)}$$

Where,

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{80}{8} = 10 \qquad \text{and} \qquad \bar{y} = \frac{1}{n}\Sigma y = \frac{648}{8} = 81$$

Also,

$$\text{Cov}(x,y,) = \frac{1}{n}\Sigma xy - \bar{x}\,\bar{y} \quad \Big| \quad V(x) = \frac{1}{n}\Sigma x^2 - \bar{x}^2 \quad \Big| \quad V(y) = \frac{1}{n}\Sigma y^2 - \bar{y}^2$$

$$= \frac{6727}{8} - 10 \times 81 \quad \Big| \quad = \frac{1018}{8} - 10^2 \quad \Big| \quad = \frac{52856}{8} - 81^2$$

$$= 840.75 - 810 \quad \Big| \quad = 127.25 - 100 \quad \Big| \quad = 6607 - 6561$$

$$\therefore \textbf{Cov(x,y)} = 30.75 \quad \Big| \quad V(x) = 27.25 \quad \Big| \quad V(y) = 46$$

Now we find,

Regression co-efficient of X on Y Regression co-efficient of X on Y

$$b_{xy} = \frac{Cov(x,y)}{V(y)} \quad \Big| \quad b_{yx} = \frac{Cov(x,y)}{V(x)}$$

$$= \frac{30.75}{46} \quad \Big| \quad = \frac{30.75}{27.25}$$

$$\therefore \; b_{xy} = 0.67 \quad \text{and} \quad \Big| \quad b_{yx} = 1.13$$

Now substituting the values of $\bar{x}$, $\bar{y}$, b_{xy} and b_{yx} in the regression equations we get,

(x-10) = 0.67(y-81) -------x on y (i)

(y-81)=1.13(x-10) ------- y on x (ii)

as the two regression equations.

Now to estimate Performance rating (y) when Experience (x) = 15, we use the regression equation of y on x

$\therefore$ (y-81) =1.13(15-10)

$\therefore$ y = 81+ 5.65 = 86.65

Hence the estimated performance rating for the operator with 15 years of experience is approximately 86.65 i.e approximately 87

10.5.2 Regression coefficients in terms of correlation coefficient.

We can also obtain the regression coefficients b_{xy} and b_{yx} from standard deviations, σ_x., σ_y and correlation coefficient 'r' using the formulas

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad \text{and} \quad b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

Also consider,

$$b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y}\, r\frac{\sigma_y}{\sigma_x} = r^2 \qquad \text{i.e. } r = \sqrt{b_{xy} \times b_{yx}}$$

Hence the correlation coefficient 'r' is the geometric mean of the regression coefficients, b_{xy} and b_{yx}

Example 5:

You are given the information about advertising expenditure and sales:

	Exp. on Advertisiment (Rs. In Lakh)	Sales (Rs. In Lakh)
Mean	10	90
S.D.	3	12

Coefficient of correlation between sales and expenditure on Advertisement is 0.8.Obtain the two regression equations.

Find the likely sales when advertisement budget is Rs. 15 Lakh.

Solution: We define the variables,

X: Expenditure on advertisement

Y: Sales achieved.

Therefore we have,

$\bar{x}$ = 10, $\bar{y}$ =90, 6x= 3, 6y = 12 and r = 0.8

Now, using the above results we can write the two regression equations as

$$(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$ -------x on y (i)

$$(y-\bar{y}) = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$ ------- y on x (ii)

Substituting the values in the equations we get,

$$(x-10) = 0.8\frac{3}{12}(y-90)$$

i.e x- 10 = 0.2 (y-90) -------x on y (i)

also $(y-90) = 0.8\frac{12}{3}(x-10)$

i.e. y-90 = 3.2 (x-10) -------y on x (ii)

Now when expenditure on advertisement (x) is 15, we can find the sales from eqn (ii) as,

y-90 = 3.2 (15-10)

∴ y = 90 + 16 = 106

Thus the likely sales are Rs.106 Lakh.

Example 6: Comput the two regression equations on the basis of the following information:

	X	Y
Mean	40	45
Standard deviation	10	9

Karl Pearson's coefficient of correlation between x and y = 0.50.

Also estimate the value of x when y = 48 using the appropriate equation.

Solution: We have,

$\bar{x}$ = 40, $\bar{y}$ =45, σx= 10, σy = 9 and r = 0.5

Now, we can write the two regression equations as

$$(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$ -------x on y (i)

$$(y-\bar{y}) = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$ ------- y on x (ii)

Substituting the values in the equations we get,

$$(x-40) = 0.5\frac{10}{9}(y-45)$$

i.e x- 40 = 0.55 (y-45) -------eqn of x on y (i)

and $(y-45) = 0.5\frac{9}{10}(x-40)$

i.e. y-45 = 0.45(x-40) -------eqn of y on x (ii)

Now when y is 48, we can find x from eqn (i) as,

x-40 = 0.55(48-45)

$\therefore$ x = 40 +1.65 = 41.65

Example 7:

Find the marks of a student in the Subject of Mathematics who have scored 65 marks in Accountancy Given,

Average marks in Mathematics	70
Accountancy	80
Standard Deviation of marks in Mathematics	8
in Accountancy	10

Coefficient of correlation between the marks of Mathematics and marks of Accountancy is 0.64.

Solution: We define the variables,

X: Marks in Mathematics

Y: Marks in Accountancy

Therefore we have,

$\bar{x} = 70,\ \bar{y} = 80,\ \sigma_x = 8,\ \sigma_y = 10$ and $r = 0.64$

Now we want to approximate the marks in Mathematics (x), we obtain the regression equation of x on y, which is given by

$$(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \qquad \text{-------x on y (i)}$$

Substituting the values we get,

$$(x-70) = 0.64\frac{8}{10}(y-80)$$

i.e x- 70 = 0.57 (y-80)

Therefore, when marks in Accountancy (Y) = 65

x- 70 = 0.57(65-80)

$\therefore$ x = 70-2.85 = 67.15 i.e. 67 appro.

Use of regression equations to find means $\bar{x}, \bar{y}$ S.D.s σ_x, σ_y and correlation coefficient 'r'

As we have that, we can obtain the regression equations from the values of Means, standard deviations and correlation coefficients 'r', we can get back these values from the regression equations.

Now, we can note that the regression equation is a linear equation in two variables x and y. Therefore, the linear equation of the type Ax+By+C = 0 or y = a+bx represents a regression equation.

e.g. 3x+5y-15 = 0 and 2x+7y+10 = 0 represent the two regression equations.

The values of means $\bar{x}, \bar{y}$ can be obtain by solving the two equations as the simultaneous equations.

Example 8:

From the following regression equation, find **means** $\bar{x}, \bar{y}$, σ_x, σ_y and 'r'

3x-2y-10 = 0, 24x-25y+145 = 0

Solution: The two regression equations are,

3x-2y-10 = 0 --------(i)

24x-25y+145 = 0 ---(ii)

Now for $\bar{x}$ and $\bar{y}$ we solve the two equations as the simultaneous equations.

Therefore, by (i) x 8 and (ii) x1, we get

24x-16y-80 = 0

24x-25y+145 = 0

\- + -

9y-225 = 0 $\qquad y = \frac{225}{9} = 25$

Putting y = 25 in eqn (i), we get

3x-2(25)-10 = 0

3x – 60 = 0 $\qquad x = \frac{60}{3} = 20$

Hence $\bar{x} = 20$ and $\bar{y} = 25$.

Now to find 'r' we express the equations in the form y=a+bx

So, from eqns (i) and (ii)

$y = \frac{3x}{2} - \frac{10}{2}$ and $y = \frac{24x}{25} + \frac{145}{25}$

$\therefore b_1 = \frac{3}{2} = 1.5$ $\qquad \therefore b_2 = \frac{24}{25} = 0.96$

Since, $\mathbf{b_1 > b_2}$ **(i.e.** $\mathbf{b_2}$ is smaller in number irrespective of sign + or -)

∴ Equation (ii) is regression of y on x and $\mathbf{b_{yx} = 0.96}$
Hence eqn (i) is regression of x on y and $b_{xy} = 1/1.5 = 0.67$

Now we find, $r = \sqrt{b_{xy} \times b_{yx}}$ i.e. $r = \sqrt{0.67 \times 0.96} = +0.84$
(The sign of 'r' is same as the sign of regression coefficients)

Example 9:
Find the means values of x,y, and r from the two regression equations. $3x+2y-26=0$ and $\sigma x+y-31=0$. Also find σ_x when $\sigma_y = 3$.

Solution: The two regression equations are,

$3x+2y-26=0$ -------- (i)
$6x+y-31=0$ ----------(ii)

Now for x and y we solve the two equations as the simultaneous equations.
Therefore, by (i) x 2 and (ii) x1, we get

$6x+4y-52 = 0$
$6x+ y-31 = 0$
$- \quad - \quad +$

$3y-21 = 0$

$\therefore y = \frac{21}{3} = 7$

Putting y = 7 in eqn (i), we get

$3x+2(7)-26 = 0$

$3x - 12 = 0$ $\qquad x = \frac{12}{3} = 4.$

Hence x = 4 and y = 7.
Now to find 'r' we express the equations in the form y=a+bx
So, from eqns (i) and (ii)

$y = -\frac{3}{2}x - \frac{26}{2}$ and $y = -\frac{6}{1}x + \frac{31}{1}$

$\therefore \mathbf{b_1 = -\frac{3}{2} = -1.5}$ $\qquad \therefore \mathbf{b_2 = \frac{6}{1} = -6}$

since, $\mathbf{b_1 < b_2}$ **(i.e.** $\mathbf{b_1}$ is smaller in number irrespective of sign + or -)

∴ Equation (i) is regression of y on x and $\mathbf{b_{yx} = -1.5}$
Hence, eqn (ii) is regression of x on y and $b_{xy} = -1/6 = -0.16$

Now we find, . r = $\sqrt{b_{xy} \text{ x } b_{yx}}$ $r = \sqrt{0.16x1.5} = = - 0.16$

Note: ***The sign of 'r' is same as the sign of regression coefficients***

Now to find 6x when 6y = 3, we use the formula,

$$b_{yx} = r\frac{\sigma_x}{\sigma_y}$$

$$\textbf{- 1.5} = -\frac{0.16x3}{6x}$$

$$\therefore \textbf{6x} = \frac{0.48}{1.5} = \textbf{0.32}$$

Hence means $\bar{x} = 4, \bar{y} = 7$, r = - 0.16 and 6x = 0.32.

EXERCISES

1. What is mean by Regression? Explain the use of regression in the statistical analysis.

2. Why are there two Regressions? Justify.

3. State the difference between Correlation and Regression.

4. Obtain the two regression equations from the data given bellow.

X: 7 4 6 5 8

Y: 6 5 9 8 2

Hence estimate y when x = 10.

5. The data given below are the years of experience (x) and monthly wages (y) for a group of workers. Obtain the two regression equations and approximate the monthly wages of a workers who have completed 15 years of service.

Experience: In years	11	7	9	5	8	6	10
Monthly wages: (in '000Rs.)	10	8	6	8	9	7	11

6. Following results are obtained for a bivariate data. Obtain the two regression equations and find y when x = 12

n = 15 $\Sigma x = 130$ $\Sigma y = 220$ $\Sigma x^2 = 2288$ $\Sigma y^2 = 5506$ $\Sigma xy = 3467$

7. Marks scored by a group of 10 students in the subjects of Maths and Stats in a class test are given below. Obtain a suitable regression equation to find the marks of a student in the subject of Stats who have scored 25 marks in Maths.

Student no:	1	2	3	4	5	6	7	8	9	10
Marks in Maths	13	18	9	6	14	10	20	28	21	16
Marks in Stats:	12	25	11	7	16	12	24	25	22	20

9. The data given below are the price and demand for a certain commodity over a period of 7 years. Find the regression equation of Price on Demand and hence obtain the most likely demand for the in the year 2008 when it's price is Rs.23.

Year:	2001	2002	2003	2004	2005	2006	2007
Price (in RS):	15	12	18	22	19	21	25
Demand (100 units)	89	86	90	105	100	110	115

10. For a bivariate data the following results were obtained

$\bar{x} = 53.2$, $\bar{y} = 27.9$, 6x = 4.8, $\sigma y = \sigma.4$ and r =0.75

Obtain the two regression equations, find the most probable value of x when y =25.

11. A sample of 50 students in a school gave the following statistics about Marks of students in Subjects of Mathematics and Science,

Subjects:	Mathematics	Science
Mean	58	79
S.D.	12	18

Coefficient of correlation between the marks in Mathematics and marks in Science is 0.8. Obtain the two regression equations and approximate the marks of a student in the subject of Mathematics whose score in Science is 65.

12. It is known that the Advertisement promotes the Sales of the company. The company's previous records give the following results.

	Expenditure on Advertisement (Rs. In Lakh)	Sales (Rs. In Lakh)
Mean	15	190
S.D.	6	20

Coefficient of correlation between sales and expenditure on Advertisement is 0.6. Using the regression equation find the likely sales when advertisement budget is Rs.25 Lakh.

12. Find the values of x,y, and r from the two regression equations given bellow. 3x+2y-26=0 and 6x+y-31=0. Also find 6x when $\sigma y = 3$.

13. Two random variables have the regression equations:

    5x+7y-22=0 and 6x+2y-20=0. Find the mean values of x and y. Also find S.D. of x when S.D. of y = 5.

14. The two regression equations for a certain data were y = x+5 and 16x = 9y-94. Find values of $\bar{x}$, $\bar{y}$ and r. Also find the S.D. of y when S.D. of x is 2.4.

11

SAMPLING DISTRIBUTION

11.0 Objectives

- To understand the relationship between two relevant characteristics of a statistical unit.
- Learn to obtain the numerical measure of the relationship between two variables.
- Use the mathematical relationship between two variables in order to estimate value of one variable from the other.
- Use the mathematical relationship to obtain the statistical constants line means and S.D.'s

11.1 Sampling Distribution

In the first chapter we have seen that there are two methods to collect a primary data: Census and Sample survey. In Sample survey we select a sizeable amount of items from the total items in consideration. Sometimes it is not possible to study the characteristics of a variable in totality. Also it may not be sensible. For example, if we are testing the quality of a food item, we cannot test every packet as it would

destroy the whole lot. Thus we select a few of them. This sample is then analysed using various statistical tools. The result is then generalized to the entire data.

Before studying about the core topic of this chapter let us get familiarized with the terminology used for this statistical process:

Population:

The totality of items for which the statistical investigation is done is called *popualation*. A Statistical population may be finite or infinite (population size).

Finite Population:

If the numbers of all possible items which are under investigation are finite, we say that the population is a finite population.

Infinite Population:

If the number of all possible items under investigation is not finite, we say that the population is infinite population.

Sample:

The subset of the population which is selected at random for the statistical investigation is called *sample.* Based on this sample the unknown measures of the population and other inference are made by investigators. A random sample is expected to be representing the important characteristics of the population. Otherwise it would not help in reaching proper conclusions.

Sample mean:

Consider a population of N number of items. A sample consisting of n number of items is randomly selected. If $x_1, x_2, x_3, \ldots\ldots, x_n$ are the items in the sample, then the *sample mean* is given by the formula: $\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots\ldots\ldots + x_n}{n}$.

Sample Proportion:

In problems related to finding the defectives, we are interested in knowing the percentage of the defectives in the sample. Such a percentage is called *sample proportion*. If d is the number of defectives in a sample of size n, then sample proportion = d/n. For example, if a coin is tossed 4000 times and a heads is observed 1800 times, then:

Sample Proportion = $\frac{1800}{4000} = 0.45$

Sample Size

The number of items in a sample is called its sample size. The size of the sample depends upon (*a*) the cost and time constraint, (*b*)the statistical behavior, *i.e.* the probability distribution of the population and sample, (*c*) information available about the measures of the population.

Sampling:

This process of selecting a sample from a given population is called *sampling*.

Parameter:

The measure of the population is called as *parameter*. For example, mean, variance, standard deviation and proportion are the popular measures used in sampling method. The value of a parameter remains the same throughout the investigation.

Statistic:

The measure of the sample is called as *statistic*. The value of a statistic varies from sample to sample. Thus, in contrast to the parameter, the statistic value is not constant.

Sampling fluctuations:

The fluctuations or variations observed in different sample statistics of the same population are termed as *sampling fluctuations.*

Sampling Distribution:

The probability distribution of the statistic (say, sample mean) is called *sampling distribution* of the statistic.

Standard Error :

In sampling we are considering measure for both population and sample at the same time. The standard deviation of the statistic is termed as S*tandard Error* (SE) just to differentiate it from the standard deviation of the population. Also, we know that the statistic value changes from sample to sample. So the sampling distribution measures these deviations, say of the sample means from the mean of all samples. Such an error is called *sampling error* or *standard error*. Students should remember that SE is nothing but the standard deviation of the sample taken for investigation.

11.2 Central Limit Theorem

This is a very important theorem in sampling distribution theory. In practice not all population show a normal distribution. The *Central limit theorem* state that "When a random sample is drawn from a population which is not normally distributed then as the sample size is increased, the standard deviation of the sample mean is approximately normally distributed with mean equal to the mean of the population and standard deviation equal to $\sigma/\sqrt{n}$, where σ is the standard deviation of the population".

If the size of the sample is more than 30 it is said to be large, in general.

11.3 Testing of Hypothesis

Hypothesis as a word means a guess or an assumed idea. In our daily life we make lot of such assumptions. In professional fields, any process starts with an assumption. The judiciary also believes that every suspect is innocent until he is proved guilty. In Statistical inferences, it is difficult to assert the population parameter like mean of the population. So a sample from the population is selected and the difference between the sample statistic and population parameter is analyzed. Such a process of analyzing the difference between the statistic and parameter is called *Testing of Hypothesis*. The conclusion of these tests is to either accept the hypothesis or to reject it.

11.4 Types of Hypothesis

11.4.1 Null Hypothesis

The hypothesis which is to be tested is called *Null Hypothesis*. It is denoted by H_0. The null hypothesis assumes that there is no difference between the statistic and the parameter. In case of the measure being the mean, if the mean of the sample is μ_0 and the population mean is μ then we write the null hypothesis as follows:

$H_0: \mu = \mu_0$ or $H_0: \mu > \mu_0$ or $H_0: \mu < \mu_0$

The latter two types will be discussed a little later in *One Tailed Tests.*

11.4.2 Alternative Hypothesis

The alternate (or opposite) of the null hypothesis is called *Alternative Hypothesis*. It is denoted by H_1.If H_0 is true then H_1 is false and vice versa.

1. If H_0: $\mu = \mu_0$, then H_1: $\mu \neq \mu_0$
2. If H_0: $\mu > \mu_0$, then H_1: $\mu \leq \mu_0$
3. If H_0: $\mu < \mu_0$, then H_1: $\mu \geq \mu_0$

11.5 Errors in making a decision

Consider a case when a college has to give a scholarship to all students whose performance has increased in the second term compared with the first term. If the college administration assumes that there is no difference in the performance (null hypothesis) then what are the decisions taken for a particular student selected at random? There are four possibilities:

1. The student's performance has improved and is given scholarship: Correct decision!
2. The student's performance has not improved but is given a scholarship. Good Luck!
3. The student's performance has not improved and is not given the scholarship. Correct decision.
4. The student's performance has not improved but is given the scholarship. Bad decision!

The remarks given are mine and need not be compared with any statistical conclusions. But roughly we come across some mistakes or errors in making decision. The decision number 2 and 4 are the wrong decisions or errors. In testing a hypothesis these errors are named as Type I and Type II error.

11.5.1 Type I Error

The error made when the null hypothesis is rejected when it is true is called *Type I error*.

Level Of Significance:

The probability that a Type I error is made is called as *level of significance*. It is denoted by α. It is the risk taken by the decision making body of making a Type I error. In majority of cases we assume a 5% level of significance for testing a hypothesis. This means that the probability rejecting the null hypothesis when it is true is only 0.05. In other words, the probability of making a correct decision is 95%. There are situations where more accuracy or less accuracy is required. In such cases the level of significance is taken as 1% or 10%.

11.5.2 Type II Error

The error made when the null hypothesis is accepted when it is not true is called *Type II error*. The probability of making such an error is denoted by β

It is for the decision maker to decide which type of error he wants to avoid. It is not possible to control both the errors at the same time, as a decrease in one type leads to an increase in the probability of the other type of error.

Power of a test:

The *power of a test* is the probability that a Type II error is avoided. The probability of making a Type II error is β, so the formula to find the power of a test is as follows:

Power of a test = 1 – β

The two types of errors can be tabulated as shown below:

	H_0 is accepted	H_0 is rejected
H_0 is true	*Correct decision*	**Type I Error**
H_0 is false	**Type II Error**	*Correct decision*

11.6 Critical Region

In testing a hypothesis we assume a normal distribution of the random variable. For drawing Statistical inferences, the standard normal variate *z* is used. We know from the chapter on probability distribution, that $z = \frac{X - \mu}{SE}$, where μ is the population mean, S.E. is the standard deviation of the sample and μ_0 is the sample mean.

From the normal table, we observe that area between *z* = –1.96 and *z* = 1.96 is 95 % of the total area. (The sum of the values is 0.4750 + 0.4750 = 0.95). Similarly, the area between *z* = –2.58 and *z* = 2.58 is 99.02% of the total area and the area between *z* = – 1.64 and

z = 1.64 is 90% of the total area. The knowledge of these areas leads us to the probabilities of the confidence level which determines the level of significance. Thus, for a 5%level of significance we know now that the *z* value should be between –1.96 and +1.96.These limits are called as *confidence limits*. The region between these confidence limits is called the *region of acceptance*. The region beyond these limits is the region of rejection called as the *critical region*.

If the value of z is beyond these confidence limits *i.e.* $z > -1.96$ or $z > 1.96$ then we conclude that it is not only due to sampling fluctuations but some more serious reasons. This is because the probability of such fluctuations is only 0.05, which is very small. Thus, we say that the difference between the statistic and parameter is highly significant. As a result the null hypothesis is rejected.

If the value of z_α is in the critical region, *i.e.* $-1.96 \leq z \leq 1.96$, then we say that the difference is observed due to some sampling fluctuations and is not significant. As a result the null hypothesis is accepted.

The same rule is followed for 1% and 10% level of significance. The confidence limits for the three levels of significance are as shown below:

1%	5%	10%
$-2.58 \leq z \leq 2.58$	$-1.96 \leq z \leq 1.96$	$-1.64 \leq z \leq 1.64$

Fig 11.1

11.7 Steps to Test a Hypothesis

The steps to test a hypothesis are as follows:

1. To decide the Null Hypothesis: It is the first step to determine, our assumption about the population parameter, which is to be tested on the basis of a sample statistic. There are three null hypotheses: $\mu = \mu_0$ or $\mu > \mu_0$ or $\mu < \mu_0$.

2. To decide the level of significance: This is the probability of making a Type I error that is of rejecting the true null hypothesis. Generally we take 5% level of significance. For quality testing of a drug 1% level of significance is considered, while for pre-poll and exit poll surveys a 10% level of significance may be selected. It is left to the concerned authority's discretion to select the level of confidence based on his requirement.

3. Critical Region: Once the level of significance is decided the critical region follows immediately. It is the region of acceptance of the null hypothesis.

4. Test Statistic: After the confidence limits are decided now we require a statistical test to analyse the sample statistic. This is called as test statistic. Since we are going to study about large sample with normal distribution the S.N.V. z is used as the test statistic.

5. Decision Making: The last and important step for testing a hypothesis is making a decision based on the result of the test statistic. The conclusion is

either to accept H_0 or to reject H_0. There are two tests for making this decision depending upon what is our H_1? We have seen before that there are three possibilities for H_1, so there are the following three tests:

a) **Two Tailed Test**: If we observe the standard normal curve we see that due to symmetry of the curve, it is moving infinitely on the sides of the mean. These two ends are called *tails* of the curve. If the null hypothesis is that $\mu = \mu_0$, then we use the two tailed test as the alternative hypothesis is $\mu \neq \mu_0$ which means either $\mu < \mu_0$ or $\mu > \mu_0$. Thus we have to check the statistic for both the tails.

For 5% level of significance, if $|z_{cal}| < 1.96$, H_0 is accepted and if $|z_{cal}| > 1.96$ H_0 is rejected.

b) **Left Tailed Test**: If our null hypothesis is H_0: $\mu = \mu_0$, and H_1: $\mu < \mu_0$. Here the area of rejection is to the left of the normal curve, hence the test is called *left tailed test*.

For 5% level of significance, if $z_{cal} \leq -1.64$, then H_0 is rejected. The value –1.64 corresponds to the 47.5% area to the left of the normal curve.

c) **Right Tailed Test**: If our null hypothesis is H_0: $\mu = \mu_0$, and H_1: $\mu > \mu_0$. Here the area of rejection is to the right of the normal curve, hence the test is called *right tailed test*.

For 5% level of significance, if $z_{cal} \geq 1.64$, then H_0 is rejected. The value 1.64 corresponds to the 47.5% area to the right of the normal curve.

The following will make the idea more clear:

Fig 12.2

11.8 Large Sample Tests

Now we shall see some examples using the *z*-test, where a large sample ($n > 30$) is taken. In problems where the sample size is small (*i.e.* $n < 30$), another test called the *t*-test or the student's *t* – test is used. We will confine ourselves to the first test.

11.8.1 Large Sample Tests for mean

Example 1:

A random sample of 100 bundles gives a mean of 8.5 tons and standard deviation 4 tons. Can the sample be regarded as drawn from a population with mean 7 tons? Test this at 5% level of significance.

Ans: Given: $\mu = 7$, $X = 8.5$, SE = 4 and $n = 100$

If the standard deviation of the population is not known, the sample standard deviation is to be taken $\therefore \sigma = SE = 4$

Null Hypothesis: $H_0: \mu = 7$ Alternative Hypothesis: $H_1: \mu \neq X$

Now, $z = \frac{X - \mu}{\sigma / \sqrt{n}} = \frac{8.5 - 7}{4 / \sqrt{100}} = -3.75$

$\therefore |z_{cal}| = 3.75$

At 5% level of significance the value of $z_\alpha = 1.96$

It is observed that $|z_{cal}| > z_\alpha$. Thus, the null hypothesis is rejected.

Thus, the sample cannot be regarded as being taken from the population with mean 7 tons

Example 2:

A machine produces copper plates of thickness 2cm with standard deviation of 0.4 cm. A sample of 50 copper plates is selected at random. The average thickness of the sample is 2.04cm. Test the hypothesis that the machine is performing in a normal way, at 5% level of significance.

Ans: Given: $\mu = 2$, $X = 2.04$, $\sigma = 0.4$ and $n = 50$

Let $H_0: \mu = X$ and $H_1: \mu \neq X$

Now, $z = \frac{X - \mu}{\sigma / \sqrt{n}} = \frac{2.04 - 2}{0.4 / \sqrt{50}} = \frac{0.04}{0.05656} = 0.71 < 1.96$

$\therefore z_{cal} = 0.71$ and at 5% level of significance, we know that $z_\alpha = 1.96$

Since $z_{cal} < z_\alpha$, we conclude that the null hypothesis is accepted.

Thus, the performance of the machine producing the copper plates is normal.

Example 3:

Uniliver Company manufacture water filters and claim that their water filters have a life of atleast 18 months. Test their claim if a sample 100 water filters taken at random had an average life of 16 months with standard deviation 6 months.

Ans: Given: $\mu = 18$, $X = 16$, $\sigma = 6$ and $n = 100$

This is an example of one tailed test. Here the null hypothesis is that the average life is atleast 18 months.

Let $H_0: \mu \geq 18$ and $H_1: \mu < 18$

Now, $z = \frac{X - \mu}{\sigma/\sqrt{n}} = \frac{16-18}{6/\sqrt{100}} = \frac{-2}{0.6} = -3.33$

$\therefore z_{cal} = -3.33$ and at 5% level of significance, we know that $z_\alpha = -1.96$

Since $z_{cal} < z_\alpha$, we conclude that the null hypothesis should be accepted.

Thus, the claim of the company is proved correct.

Example 4:

A pay commission is appointed to study the wages of government employees. It was provided with the information that the average salaries of the employees are Rs. 8,400 with standard deviation Rs. 3000. But the commission selected 100 employees at random and found that their average salary is Rs. 8,800. Test at 5% level of significance, whether the sample chosen is a representative of the population?

Ans: Given: $\mu = 8400$, $X = 8800$, $\sigma = 3000$ and $n = 100$

Let $H_0: \mu = X$ and $H_1: \mu \neq X$

Now, $z = \frac{X - \mu}{\sigma/\sqrt{n}} = \frac{8800-8400}{3000/\sqrt{100}} = \frac{400}{300} = 1.33 < 1.96$

$\therefore z_{cal} = 1.33$ and at 5% level of significance, we know that $z_\alpha = 1.96$

Since $z_{cal} < z_\alpha$, we conclude that the null hypothesis is accepted.

The sample chosen by the commission represents the population of employees.

11.8.2 Difference between means

If two samples of size n_1 and n_2 are drawn from a population with means μ_1, μ_2 and standard deviations σ_1, σ_2 respectively then the sampling distribution of the

difference between the sample means X_1 and X_2 follows a normal distribution with mean $\mu_1 - \mu_2$, standard error SE = $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ and $z = \frac{X_1 - X_2}{SE}$

In problems of this kind we first find the standard error and then the test statistic z.

Example 5:

The average income of 100 men in a city is Rs. 15,000 with standard deviation Rs. 8,500 and the average income of 100 women is Rs. 12,000 and standard deviation Rs. 9000. Can it be said at 5% level of confidence that there is a significant difference between the average income of men and women?

Ans: H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$

Given: For men: $n_1 = 100$, $X_1 = 15000$, $\sigma = 8500$

For women: $n_2 = 100$, $X_2 = 12000$, $\sigma = 9000$

$$\text{SE} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(8500)^2}{100} + \frac{(9000)^2}{100}} = \sqrt{722500 + 810000}$$

$$\therefore \text{SE} = 1237.94$$

$$\therefore z = \frac{X_1 - X_2}{SE} = \frac{15000 - 12000}{1237.94} = 2.42 > 1.96$$

At 5% level of significance the value of $z_\alpha = 1.96$

It is observed that $z_{cal} > z_\alpha$. Thus, the null hypothesis is rejected.

Hence there is a significant difference between the salaries of men and women.

11.8.3 Large Sample Tests for Proportion

In situations when the population and sample is expressed in percentages or proportions, the method of testing the hypothesis is as follows:

If the population proportion is π with standard deviation σ and the sample proportion is p with standard error SE = $\sqrt{pq/n}$, then the test statistic is $z = \frac{\pi - p}{SE}$. In problem for testing the hypothesis fro proportion we first find the standard error and then the test statistic.

Example 6:

A manufacturer claims that 10% of his product is defective. A sample of 300 items selected at random had 32 defective items. Test his claim at 1% level of significance.

Ans: H_0: $\pi = 10\% = 0.1$ and H_1: $\pi \neq 0.1$

Given: $\pi = 10\% = 0.1$, $p = \frac{32}{300} = 0.11 \quad \Rightarrow q = 1 - p = 0.89$

$$\therefore \text{SE} = \sqrt{pq/n} = \sqrt{\frac{0.11 \text{ x } 0.89}{400}} = 0.016$$

$$\therefore z = \frac{\pi - p}{SE} = \frac{0.1 - 0.11}{0.016} = -0.625$$

At 1% level of confidence the value of $|z_\alpha| = 2.58$

$\therefore |z_{cal}| = 0.626 < 2.58$

Thus, the null hypothesis is accepted. Hence the manufacturer's claim is accepted.

Example 7:

A die is thrown 5000 times and a throw of 2 or 6 is observed 1520 times. Test whether the die is biased?

Ans: Let H_0: The die is unbiased and H_1: The die is biased

Given: $\pi = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} = 0.33$ (since the probability of getting a 2 or 6 is 1/6)

$$p = \frac{1520}{5000} = 0.304 \Rightarrow q = 1 - p = 0.696$$

$$\therefore \text{SE} = \sqrt{pq/n} = \sqrt{\frac{0.304 \text{ x } 0.696}{5000}} = 0.006$$

$$\therefore z = \frac{\pi - p}{SE} = \frac{0.33 - 0.304}{0.006} = 4.33 > 1.96$$

At 5% level of confidence the value of $z_\alpha = 1.96$

$\therefore z_{cal} > z_\alpha$, we reject the null hypothesis and conclude that the die is biased.

11.8.4 For Difference between proportions

Let two random samples of size n_1 and n_2 with proportions p_1 and p_2 respectively have the standard error SE = $\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$, where $p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$.

The test statistic is then given by $z = \frac{p_1 - p_2}{SE}$. We first find the combined proportion p and compute the SE. Then the test statistic is calculated.

Example 8:

An old machine produced 10 defective bolts in a batch of 300. After the servicing was done the same machine was found to produce 6 defective bolts in a batch of 200. Help the manufacturer to conclude whether the machine has improved after the servicing?

Ans: Let H_0: $p_1 = p_2$ and H_1: $p_1 \neq p_2$

Given: $p_1 = \frac{10}{300} = 0.033$ and $p_2 = \frac{6}{200} = 0.03$

$n_1 = 300$ and $n_2 = 200$

$\therefore p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{10+6}{500} = 0.032 \quad \Rightarrow q = 1 - p = 0.968$

$\therefore \text{SE} = \sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{0.032 \text{ x } 0.968 \text{ x } \left(\frac{1}{300}+\frac{1}{200}\right)} = \sqrt{0.0309 \text{ x } \left(\frac{500}{60000}\right)}$

$\therefore \text{SE} = 0.016$

$\therefore z = \frac{p_1 - p_2}{SE} = \frac{0.033 - 0.03}{0.016} = 0.1875 < 1.96$

At 5% level of confidence, we observe that $z_{cal} < z_\alpha$.

Thus, the null hypothesis is accepted.

The machine has not improved.

11.9 Interval Estimation

We know from the discussion till now the different level of significance and their confidence limits. For 5% level of significance the limits for z are ± 1.96. This can be written as $-1.96 \le \frac{X-\mu}{SE} \le 1.96 \quad \Rightarrow \quad -1.96\ SE \le X - \mu \le 1.96\ SE$

$\Rightarrow \mu - 1.96\ SE \le X \le \mu + 1.96\ SE$

$\Rightarrow$ the confidence interval for sample at 5% level of confidence is (μ –1.96SE , μ + 1.96SE)

Similarly we can derive the *confidence intervals* for different levels of confidence.

The following table demonstrates these formulae:

	Level of Significance			
	1%	5%	10%	
Sample mean	$\mu \pm 2.58SE$	$\mu \pm 1.96\ SE$	$\mu \pm 1.645SE$	Here, SE = $\frac{\sigma}{\sqrt{n}}$
Population Mean	$X \pm 2.58SE$	$X \pm 1.96\ SE$	$X \pm 1.645SE$	
Sample proportion	$\pi \pm 2.58SE$	$\pi \pm 1.96\ SE$	$\pi \pm 1.645SE$	SE = $\frac{\sqrt{pq}}{n}$
Population proportion	$p \pm 2.58SE$	$p \pm 1.96\ SE$	$p \pm 1.645SE$	SE= $\frac{\sqrt{\pi(1-\pi)}}{n}$

Example 9:

A coin was tossed 200 times and heads was observed 105 times. Compute the confidence intervals at 5% level of significance.

Ans: The confidence interval for sample proportion at 5% level of significance is $\pi \pm 1.96$ SE, where SE = $\frac{\sqrt{pq}}{n}$

Now, $\pi = 0.5$ (the probability of getting heads is 1/2)

$p = \frac{105}{200} = 0.525 \Rightarrow q = 1 - p = 0.475$ and $n = 200$

$$SE = \sqrt{\frac{0.525 \text{ x } 0.475}{200}} = 0.035$$

$\therefore$ the confidence interval for sample proportion is $0.5 \pm (1.96 \text{ x } 0.035) = 0.5 \pm 0.069$

Thus, the confidence interval is (0.431, 0.569).

11.10 Determination of Sample Size

If the confidence level for a testing of hypothesis is known and the maximum error allowed ($E = X - \mu$) is given along with the standard deviation (σ) of the population then the size of the sample (n) can be determined by the formula: $n = \left(\frac{\sigma z}{E}\right)^2$

If the confidence level for testing a hypothesis is known, the maximum error allowed ($E = \pi - p$) along with the population proportion is given then the sample size (n) is calculated by the formula: $n = \frac{pq\, z^2}{E^2}$

Example 10:

The students of College of Engineering, Nagpur have designed a robot. The time this robot takes to react after a command is given has a standard deviation of 0.8sec. How large a sample of measuring the time should be taken by the students to be 95% confident of not exceeding an error of 0.1 sec?

Ans: Given: $\sigma = 0.8$, $E = 0.1$ and $z = 1.96$ (for 95% confidence level)

$$\therefore n = \left(\frac{\sigma z}{E}\right)^2 = \left(\frac{0.8 \text{ x } 1.96}{0.1}\right)^2 = 245.86 \approx 246$$

Thus, the sample size that the students should take for measurements is 246.

Exercise

(1) Define the following terms with suitable examples:

(a) Population, (b) Sample, (c) Sampling Distribution, (d) Null Hypothesis, (e) Level of Significance, (f) Level of Confidence, (g) Standard Error.

(2) Writes a short note on types of Hypothesis.

(3) Describe the steps of testing a hypothesis.

(4) Explain briefly with examples, the two types of errors.

(5) Discuss the two tailed test and one tailed test with suitable diagram.

(6) State the Central Limit Theorem.

(7) A random sample of 225 observations of an item had mean 45 and standard deviation 12. At 5% level of significance can you conclude that the sample is derived from a population of mean 42.

(8) The average marks of students in a college are 58 with standard deviation 16. A random sample of 100 students selected had average marks 61. Test that the sample is representative of the class at 5% level of significance.

(9) A building contractor for Regency Builders claims that the average wages paid to the workers are Rs. 100 per day with standard deviation 9. The Proprietor takes a sample of 80 workers and finds their average wages as Rs. 85. Should he believe the contractor? Test the claim of the contractor at 5% level of significance.

(10) A random sample of 100 cakes had average weight 80 gm and standard deviation 32 gm. If the population weight is 78 gm, test the hypothesis that the sample is taken from the same population.

(11) The average IQ of children in a village is 55. A sample of 200 children selected at random from the village gave the average IQ as 40 with standard deviation 25. Can you conclude that the sample represents the population?

(12) A sample of size 400 was drawn and the sample mean was found to be 99. Test whether this sample could have come from a normal population of mean 100 and standard deviation 8?

(13) The manufacturers of Speed Petrol claim that the mileage of a four wheeler using their fuel is 22km/hr with standard deviation of 6km/hr. 100 cars are tried at random with one litre speed petrol. The sample average mileage of a car was 21km/hr. Test the claim of the company at 1% level of significance.

(14) The average life of a chain smoker is estimated by a government agency as 46 years. A sample 250 chain smokers were studied and it was found that their average life was 40 years with standard deviation 16 years. Test the claim of the government agency.

(15) The average life of a tyre of a branded car is claimed to be 3600 km with standard deviation 360 km.. 100 samples from a rural area were taken at random. It was found that the average life of a tyre in that area is 3000 km. Can you conclude from this that the tyres of the branded car have a less life in rural area?

(16) A random sample of 50 items gives the mean 6.2 and the standard deviation 10.24. Can it be regarded as drawn from a normal population with mean 5.4 at 5% level of significance?

(17) A sample of 100 tyres is taken from a lot. The mean life of the tyres are found to be 39, 350 km with a standard deviation of 3260 km. Could the sample come from a population with mean life 40,000 km? Test the hypothesis at 1%level of confidence.

(18) The average marks of a sample of 100 students in a college in the subject of Mathematics are 78 with standard deviation 36 while another sample of 120 students showed the average marks in English as 56 with standard deviation as 25. Can you conclude that there is a significant relation between the marks of the students in both the subjects?

(19) An examiner claims that his paper assessment is 90% right. A moderator takes a random sample of 100 papers and finds 12 papers nor assessed properly. Test the claim of the examiner at 5% level of confidence.

(20) The average life of an Indian is greater 70 years. A random sample of 100 Indians has an average life of 71.8 years with standard deviation of 7.8 years. Test the hypothesis.

(21) A simple sample of the heights of 6400 Englishmen has a mean of 67.85 inches and a standard deviation of 2,56 inches, while a sample of heights of 1600 Austrians has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the Austrians on an average are taller than the Englishmen?

(22) In a sample of 400 parts manufactured by a factory, the number of defective parts was found to be 30. The company, however claimed that only 5% of their product is defective. Is the claim tenable?

(23) A wholesaler claims that only 5% of the products supplied by him are defective. A random sample of 600 products contained 36 defective products. Test the claim of the wholesaler.

(24) The political advisor of a party informs his leaders that 80% of the population is in their favor for the next elections. The party leaders tell an agency to make a survey and submit its report. In the survey it was found that out of 1000 people interviewed 780 people supported the party. Test the claim of the political advisor.

(25) 600 out of 1350 people in a village A watch TV serials while 550 out of 1060 people in village B watch TV serials. Can you conclude from this data that there is a significant difference between the tastes of the villagers of both the villages?

(26) A coin is tossed 700 times and the heads appeared 360 times. Can you conclude that the coin is unbiased?

(27) In a sample of 350 people in Mumbai, 180 were non vegetarians. Can we conclude from this that there are equal percentage of vegetarians and Non-vegetarians in Mumbai?

(28) In an examination 100 students of College A got average marks 67 with standard deviation 6 and that of 200 students of College B got average marks 76 with standard deviation 9. Is there a significant difference between the performance of the students of College A and B? Test the hypothesis at 90% and 95% level of confidence.

(29) The average pay of 60 men in a factory is Rs. 120 with standard deviation Rs. 24 and that of women is Rs. 90 with standard deviation Rs. 18. Is there a significant difference between the payment to men and women in the factory?

(30) A die is thrown 6000 times and a throw of 1 or 6 was observed to be 2340 times. Can you conclude that the die is unbiased?

(31) From Versova, a sample of 100 fish is selected at random and its length is measured to be 21 cm with standard deviation 13 cm. From Vasai, another sample of 150 fish is selected and their average length is measured to be 18 cm with standard deviation 10.2 cm. Is there a significant difference between the two samples of fish taken from two markets?

(32) 300 out of 550 people in a survey were men and 220 out of 400 were found to be men in an another survey. Does these survey represent the same population?

(33) In a random sample of 600 men taken from a city, 350 are found to be smokers. In another random sample of 800 men, 400 are found to be smokers.

Do the data indicate that there is significant difference between the habit of smoking in the two cities?

(34) A slim converter is to be launched by UCA ltd for people who are overweight. The product is for people whose weight is more than 75 kg. The company does a social survey of 1200 people and finds that 260 people are overweight. The company will launch the product in the market only if it is sure that at least 20% of the people are overweight. Determine the hypothesis and test it for 95% confidence level.

(35) Cardiac patients were implanted pacemakers. A plastic connector module mounts on the top of the pacemaker. Assuming that the standard deviation of 0.0015 inches. Find the 95% confidence level interval for the mean size of the connector.

(36) Given a population with a standard deviation of 8.6. What is the sample size needed to estimate the mean of population with ± 0.5 with 99% confidence?

(37) What should be the sampling size at 95% level of confidence with maximum possible error is 4 and the standard deviation is 36?

(38) A coin was tossed 2000 times and heads was observed 1020 times.
Compute the confidence intervals at 5% level of significance.

(39) The population standard deviation is 38. A random sample of 100 observations gave the mean as 120. Find the 99% confidence intervals for the population mean.

(40) What should be the sample size taken by a exit poll agency from a constituency if they want to maintain 90% confidence level with the population proportion 0.2 and accuracy of ±0.01?

❖❖❖

UNIT 6

12

CHI-SQUARE (χ^2) DISTRIBUTION AND ITS PROPERTIES

Unit Structure

12.1 Objectives

- After going through this chapter students will be able to understand:
- The Chi-square distribution.
- The Chi-square test statistic.
- Uses of the Chi-square test.
- Pair of categorical variables can be summarized using contingency table.
- Perform a Chi-square goodness of fit test.
- The Chi-square test can compare an observed contingency table to an expected table and determine if the categorical variable are independent.
- YATE'S Correction for Contingency table.

12.2 Introduction

Chi-square (χ^2) test is a nonparametric statistical analyzing method often used in experimental work where the data consist in frequencies or 'counts' – for example the number of boys and girls in a class having their tonsils out – as distinct from quantitative data obtained from measurement of continuous variables such as temperature, height, and so on. The most common use of the test is to assess the probability of association or independence of facts.

A common problem in applied machine learning is determined whether input features are relevant to the outcome to be predicated. This is the problem of feature selection.

In the case of classification problems where input variables are also categorical, we can use statistical tests to determine whether the output variable is dependent or independent of the input variables. If independent then the input variable is a candidate for a feature that may be irrelevant to the problem and removed from the dataset.

The Pearson's Chi-square statistical hypothesis is an example of a test for independence between categorical variables.

In this chapter, you will learn the Chi-square statistical hypothesis test for quantifying the independence of pairs of categorical variables.

12.3 Chi-Square distribution

We have been discussing the distribution of mean obtained from all possible samples or a large number of samples drawn from a normal population, distribution with mean μ and variance σ^2/n. Now we are interested in knowing the distribution of sample variances s^2 of these samples. Consider a random sample $X_1, X_2, \ldots\ldots, X_n$ of size n. Let the observations of this sample be denoted by $x_1, x_2, \ldots\ldots, x_n$. We know that the variance,

$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$ for $i = 1,2, \ldots\ldots, n$

Or $\sum_i (x_i - \bar{x})^2 = (n-1)s^2 = ks^2$ where $k = (n-1)$.

A quantity ks^2/σ^2 , which is a pure number is defined as χ_k^2 . Now we will give the distribution of the random variable χ_k^2 , which was first discovered by Helmert in 1876 and later independently given by Karl Pearson in 1900. Another way to

understand Chi-square is : if $X_1, X_2, \ldots\ldots, X_n$ are n independent normal variates with mean zero and variance unity, the sum of squares of these variates is distributed as Chi-square with n degree of freedom. The Chi-square distribution was discovered mainly as a measure of goodness of fit in case of frequency distribution, i.e. whether the observed frequencies follow a postulated distribution or not. The probability density function of χ^2- variate is,

$$f_k(\chi^2) = \frac{1}{2^{k/2}\,\Gamma(k/2)}\,(\chi^2)^{\frac{1}{2}k-1}\,e^{\frac{-1}{2}\chi^2}$$

12.4 Properties of Chi-Square Distribution

The χ^2 distribution follow the following properties:

1. The whole χ^2 distribution curve lies in the first quadrant since the range of χ^2 is from 0 to ∞.

2. The χ^2 distribution has only one parameter k, the degree of freedom for χ^2. Thus, the shape of the probability density curve mainly depends on the parameter k.

3. χ^2 distribution curve is highly positive skewed.

4. It is an unimodal curve and its Mode is at the point $\chi^2 = (k-1)$.

5. The shape of the curve varies immensely especially when k is small. For k =1 and k = 2, it is just like a hyperbola.

6. χ^2 distribution is completely defined by one parameter 'k', which is known as the degree of freedom for χ^2 distribution.

7. The constants for χ^2 distributions are as follows:

Mean = $\mu = k$

Variance =$\sigma^2 = 2k$

Skewness = $\alpha_1 = 2\left(\frac{2}{k}\right)^{\frac{1}{2}}$

8. The movement generating function for χ^2 distribution is

$$\emptyset_{\chi^2}(\mathrm{t}) = (1-2t)^{-k/2}.$$

9. r^{th} raw moment of χ^2 distribution is $\mu'_r = \frac{2\Gamma\left(\frac{k}{2}+r\right)}{\Gamma k/_2}$.

10. For large degrees of freedom say $k \geq 100$, the variable is distributed normally with mean 0 and variance 1.

12.5 Test of Goodness of Fit

Generally the population study has been taken to follow a known distribution such as normal, binomial or Poisson distribution. To assume that the population is distributed normally is a common practice and hence we explain the test of goodness of fit of normal population first.

This is very powerful test for testing difference between observed data and theoretical expectation. The test is given by Karl-Pearsons and is known as χ^2 test of goodness of fit.

The test statistic used for χ^2 distribution is based on two types of frequencies namely observed frequencies denoted by O_i and expected frequencies denoted by E_i. The larger the deviation from the null hypothesis, the larger the difference between observed and expected frequencies then Karl-Pearson's χ^2 is given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Follow χ^2 distribution with $(k - p - 1)$ degree of freedom (d.f).

If the calculated value of χ^2 is greater than the table value of χ^2 for $(k - p - 1)$ d.f. and level of significance α, reject H_0. Rejection of H_0 means that the postulated theoretical distribution is not fit to the observed data, or in other words the data do not support the assertion about the theoretical distribution.

Procedure for test significance and goodness of fit:

Set up a null hypothesis and calculate

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Find the degree of freedom (d.f.) and read the corresponding value of χ^2 at a prescribed significance level from table.

From of χ^2 table we can also find the probability P corresponding to the calculated values of χ^2 for the given d.f.

If P < 0.05 the observed value of χ^2 is significant at 5% level of significance.

If P > 0.05 it is good fit and the value is not significant.

12.6 Contingency table

The data are often based on counting of objects of units. These numbers fall in various categories of attributes in a two-way classification and are very well displayed systematically in a table know as contingency table. We can express a contingency table as a rectangular array of order $(m \times n)$, having mn cells, where m denotes the number of rows which are equal to the number of categories of an attribute or criterion X and n denotes the number of columns equal to the number of categories of an attribute or criterion Y.

Attribute X	Attribute Y	Total
	Y_1 Y_2 ... ... Y_n	
X_1 X_2. . . . X_m		
Total		

2×2 Contingency table:

When a contingency table is of order 2×2, test of independence of factors can be performed in the same manner as for $(m \times n)$ contingency table. But in this situation the value of χ^2 can also be calculated directly from the observed frequencies. It is nothing but a short-cut method to obtaining the calculated value of χ^2. Suppose the contingency table of order 2×2 for two factor X and Y is as presented below.

Factor X	Factor Y		Total
	Y_1	Y_2	
X_1	a	b	(a + b)
X_2.	c	d	(c + d)
Total	(a +c)	(b+d)	a + b + c + d = n

12.7 Test of Independence of Factors

It is apparent now, that a contingency table is a rectangular array having rows and columns associated with different factors. The hypothesis that one factor is independent of the other or not, i.e. H_0 : Two factors are independent of each other.

H_1 : Two factors are dependent of each other.

The test statistic used for χ^2 distribution is based on two types of frequencies namely observed frequencies denoted by O_i and expected frequencies denoted by E_i. The larger the deviation from the null hypothesis, the larger the difference between observed and expected is.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Squaring the differences makes them all positive. Each difference is divided by the expected number, and these standardized ratios are summed: the more differences between what you would expect and what you get the bigger the number.

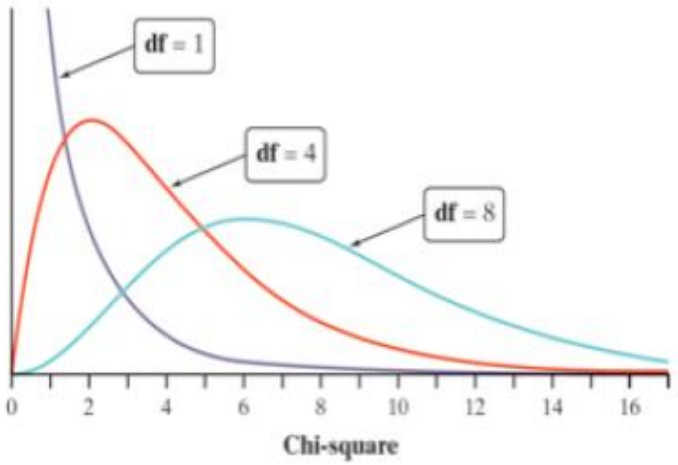
Degrees of Freedom: A critical factor in using the chi-square test is the "degrees of freedom", which is essentially the number of independent random variables involved. For example, you do a cross and see 290 purple flowers and 110 white flowers in the offspring.

Degrees of freedom is simply the number of classes of offspring minus 1.

For our example, there are 2 classes of offspring: purple and white. Thus, degrees of freedom (d.f.) = 2 -1 = 1.

Number of degrees of freedom = (number of rows -1) (number of columns – 1)

i.e. $d.f = (r-1)(c-1)$



The image above shows that the distribution of the Chi-square statistic starts at zero and can only have positive values.

The shape of the distribution is much different than the t or z statistic and is skewed to the right.

The shape of the distribution changes as the degree of freedom increses.

Critical Chi-Square: Critical values for chi-square are found on tables, sorted by degrees of freedom and probability levels. Be sure to use p = 0.05.

If your calculated chi-square value is greater than the critical value from the table, you "reject the null hypothesis". If your chi-square value is less than the critical value, you "fail to reject" the null hypothesis (that is, you accept that your genetic theory about the expected ratio is correct).

Critical values of the χ^2 Distribution table

d.f.	α								
	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801

Note:

χ^2 test is non-parametric test.

Use Nonparametric Tests:

Used when either the dependent or independent variable is ordinal.

Used when the sample size is small.

Used when underlying population is not normal.

The value of χ^2 test statistic can never be negative.

Example 1: In experiments on pea breading the following frequencies of seeds were obtained:

Round and yellow	Wrinkled and yellow	Round and green	Wrinkled and green	Total
315	101	108	32	556

Theory predicate that the frequencies should be in proportions 9:3:3:1. Examine the correspondence between theory and experiment.

Solution: First select null hypothesis $H_0 =$ The correspondence between theory and experiment.

Theory predicate that the frequencies should be in proportions 9:3:3:1.

Total value = 9 + 3 + 3 + 1= 16

O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
315	$\frac{9}{16} \times 556 = 313$	4	0.0128
101	$\frac{3}{16} \times 556 = 104$	9	0.0865
108	$\frac{3}{16} \times 556 = 104$	16	0.1538
32	$\frac{1}{16} \times 556 = 35$	9	0.2571
		Total	0.5102

The test statistic for χ^2 distribution is

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 0.5102$$

Number of degrees of freedom $= (k - 1) = (4 - 1) = 3$

χ^2 table value for 3 d.f at $\alpha = 0.05$ = 7.815.

$\Rightarrow \chi^2_{cal} < \chi^2_{tab}$

Therefore, H_0 is accepted.

Hence there is a very high degree of agreement between theory and experiment.

Example 2: A set of five similar coins is tossed 320 times and the result is

No. of heads	0	1	2	3	4	5
Frequency	6	27	72	112	71	32

Test the hypothesis that the data follow a binomial distribution.

Solution: First select null hypothesis, $H_0 =$ The data follow a binomial distribution.

P: probability of getting a head $= \frac{1}{2}$

q: probability of not getting a head $= \frac{1}{2}$

Here for expected frequencies,

P(zero head) $= {}^5C_0\, p^0 q^5 \times 320 = \frac{1}{32} \times 320 = 10$

P(one head) $= {}^5C_1\, p^1 q^4 \times 320 = \frac{5}{32} \times 320 = 50$

P(Two head) $= {}^5C_2\, p^2 q^3 \times 320 = \frac{10}{32} \times 320 = 100$

P(Three head) $= {}^5C_3\, p^3 q^2 \times 320 = \frac{10}{32} \times 320 = 100$

P(Four head) $= {}^5C_4\, p^4 q^1 \times 320 = \frac{5}{32} \times 320 = 50$

P(Five head) $= {}^5C_5\, p^5 q^0 \times 320 = \frac{1}{32} \times 320 = 10$

O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
6	10	16	1.6
27	50	529	10.58
72	100	784	7.84
112	100	144	1.44
71	50	441	8.82
32	10	484	48.4
		Total	78.68

The test statistic for χ^2 distribution is

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 78.68$$

Number of degrees of freedom $= (k-1) = (6-1) = 5$

χ^2 table value for 3 d.f at $\alpha = 0.05$ = 11.07.

$\Longrightarrow \chi^2_{cal} > \chi^2_{tab}$

Therefore, H_0 is rejected.

Hence the data follow the binomial distribution is rejected.

Example 3: Fit a Poisson distribution to the following data and test for its goodness of fit at level of significance 0.05.

X	0	1	2	3	4
F	419	352	154	56	19

Solution: First select null hypothesis, $H_0 =$ The data follow a Poisson distribution.

For Poisson distribution we need to find mean of the data.

X	F	fx
0	419	0
1	352	352
2	154	308
3	56	168
4	19	76
Total	1000	904

$$m = \frac{\sum fx}{\sum f} = \frac{904}{1000} = 0.904$$

$$\therefore\ e^{-0.904} = 0.4049$$

Here for expected frequencies are $\frac{e^{-0.904} \times 0.904^x}{x!} \times 1000$.

P(x = 0)= $\frac{e^{-0.904} \times 0.904^x}{x!} \times 1000 = 404.9$

P(x = 1)= $\frac{e^{-0.904} \times 0.904^x}{x!} \times 1000 = 366$

P(x = 2)= $\frac{e^{-0.904} \times 0.904^x}{x!} \times 1000 = 165.4$

P(x = 3)= $\frac{e^{-0.904}\times 0.904^{x}}{x!} \times 1000 = 49.8$

P(x = 4)= $\frac{e^{-0.904}\times 0.904^{x}}{x!} \times 1000 = 11.3$

In order that the total observed and expected frequencies may agree, we take the first and last theoretical frequencies as 406.2 and 12.6 instead of 404.9 and 11.3 as shown in table.

Therefore expected frequencies distribution

X	0	1	2	3	4	Total
F	404.9	366	165.4	49.8	11.3	997.4
Instead	406.2				12.6	1000

O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
419	406.2	163.84	0.403
352	366	196	0.536
154	165.4	129.96	0.786
56	49.8	38.44	0.772
19	12.6	40.96	3.251
		Total	5.748

The test statistic for χ^2 distribution is

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 5.748$$

Since the mean of the theoretical distribution has been estimated from the given data and the totals have been made to agree, there are two constrains so that the number of degree of freedom

Number of degrees of freedom $= (k - 2) = 5 - 2 = 3$

χ^2 table value for 3 d.f at $\alpha = 0.05$ = 7.815.

$\Rightarrow \chi^2_{cal} < \chi^2_{tab}$

Therefore, H_0 is accepted.

Hence, the Poisson distribution can be fitted to the data.

Example 4: A company director is concerned that his company's share may be unevenly distributed throughout the country, in a survey in which sample of 200 customers are selected from four zones and are tabulated as under:

	Zone				
	I	II	III	IV	
Purchase the brand	80	110	90	100	380
Not purchase the brand	120	90	110	100	420
	200	200	200	200	800

At $p = 0.05$, use χ^2 to determine whether the company share is same across the four zones.

Solution: First select null hypothesis H_0 and Alternative hypothesis H_1.

H_0 : The company share is same across four zones.

H_1 : The company share is not same across four zones.

At $\alpha = 0.05$.

Now calculate expected frequencies for given table:

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{380 \times 200}{800} = 95$$

$$E_{12} = \frac{R_1 \times C_2}{N} = \frac{380 \times 200}{800} = 95$$

$$E_{13} = \frac{R_1 \times C_3}{N} = \frac{380 \times 200}{800} = 95$$

$$E_{14} = \frac{R_1 \times C_4}{N} = \frac{380 \times 200}{800} = 95$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{420 \times 200}{800} = 105$$

$$E_{22} = \frac{R_2 \times C_2}{N} = \frac{420 \times 200}{800} = 105$$

$$E_{23} = \frac{R_2 \times C_3}{N} = \frac{420 \times 200}{800} = 105$$

$$E_{24} = \frac{R_2 \times C_4}{N} = \frac{420 \times 200}{800} = 105$$

O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
80	95	225	2.368
110	95	225	2.368
90	95	25	0.263
100	95	25	0.263
120	105	225	2.143
90	105	225	2.143
110	105	25	0.238
100	105	25	0.238
		Total	10.024

The test statistic for χ^2 distribution is

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 10.024$$

Number of degrees of freedom $= (r-1)(c-1) = (2-1)(4-1) = 3$

χ^2 table value for 3 d.f at $\alpha = 0.05$ = 7.815.

$\Longrightarrow \chi^2_{cal} > \chi^2_{tab}$

Therefore, H_0 is rejected.

Hence the company share is not same across four zones.

Example 5: From the following table showing the number of plants having certain characters, test the hypothesis that the flower colour is independent of flatness of leaf at the 0.1 level of significance.

	Flat leaves	Curled leaves	Total
White Flowers	99	36	135
Red Flowers	20	5	25
Total	119	41	160

Solution: First select null hypothesis H_0 and Alternative hypothesis H_1.

H_0 : The flower colour is independent of flatness of leaf.

H_1 : The flower colour is dependent of flatness of leaf.

At $\alpha = 0.1$.

Now calculate expected frequencies for given table:

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{135 \times 119}{160} = 100$$

$$E_{12} = \frac{R_1 \times C_2}{N} = \frac{135 \times 41}{160} = 35$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{25 \times 119}{160} = 19$$

$$E_{22} = \frac{R_2 \times C_2}{N} = \frac{25 \times 41}{160} = 6$$

O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
99	100	1	0.01
36	35	1	0.0286
20	19	1	0.0526
5	6	1	0.1667
		Total	0.2579

The test statistic for χ^2 distribution is

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 0.2579$$

Number of degrees of freedom $= (r-1)(c-1) = (2-1)(2-1) = 1$

χ^2 table value for 1 d.f at $\alpha = 0.1$ is 0.0185

$\Rightarrow \chi^2_{cal} > \chi^2_{tab}$

Therefore, H_0 is rejected.

Hence The flower colour is dependent of flatness of leaf.

12.8 YATES Correction

We know that the χ^2- distribution is a continuous distribution. It has been proved that if any of the cell frequency in contingency table of order 2×2 is less than 5, the continuity of χ^2- distribution curve is not maintained. So to remove this discrepancy, Yate's suggested a correction which is extensively used. He suggested that add 0.5 in the frequency which is less than 5, and subtract and add 0.5 to the

remaining cell frequencies in such a way that the marginal totals remain the same. Then calculate the value of χ^2 by formula

$$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Using adjusted contingency table.

Instead of adjusting the contingency table of order 2×2, the above formula has been amended and this takes care of the correction. The value of χ^2 under correction can directly be calculated by the formula,

$$\chi^2 = \frac{n\left(|ad-bc| - \frac{n}{2}\right)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Here $|ad-bc|$ means that we consider only the absolute value of the difference. Moreover, if one does not use the formula given specifically for the 2×2 contingency table but follows the general procedure, under Yate's correction, the formula for the χ^2 is,

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(|O_{ij} - E_{ij}| - \frac{1}{2}\right)^2}{E_{ij}}$$

It is worthwhile to point out that all three approaches yield the same value of the χ^2 because they are fundamentally the same.

Example 6: Use Yate's correction and test whether A and B are independent. Observed frequencies are as under:

	A	Not A	Total
B	45	55	100
Not B	60	40	100
Total	105	95	200

Solution:

H_0: Two attribute A and B are independent.

To use χ^2 test with Yate's correction, we have

$$n = 200, \quad a = 45, \quad b = 55, \quad c = 60, \quad d = 40$$

The χ^2 test statistic is

$$\chi^2 = \frac{n\left(|ad - bc| - \frac{n}{2}\right)^2}{(a+c)(b+d)(a+b)(c+d)}$$

$$\chi^2 = \frac{200\left(|45 \times 40 - 55 \times 60| - \frac{200}{2}\right)^2}{(105)(95)(100)(100)}$$

$$\chi^2 = \frac{200\left(|1800 - 3300| - \frac{200}{2}\right)^2}{(105)(95)(100)(100)}$$

$$\chi^2 = \frac{200(1500 - 100)^2}{(105)(95)(100)(100)}$$

$$\chi^2 = \frac{200 \times 1400 \times 1400}{(105)(95)(100)(100)} = 3.9298$$

Number of degrees of freedom $= (r-1)(c-1) = (2-1)(2-1) = 1$

χ^2 table value for 1 d.f at $\alpha = 0.05$ is 3.84

$\Longrightarrow \chi^2_{cal} > \chi^2_{tab}$

Therefore, H_0 is rejected.

Hence A and B are dependent.

Example 7: the number of licensor companies classified by the proportion of foreign profits derived from license agreements were as follows:

Proportion of profits	Licensor type		Total
	Dominant product	Diversified	
Less than 5%	1	6	7
5% or more	7	6	13
Total	8	12	20

Solution : Let H_0: Proportion of profit and licensor types are independent.

To use χ^2 test with Yate's correction, we have

$$n = 20, \qquad a = 1, \qquad b = 6, \qquad c = 7, \qquad d = 6$$

The χ^2 test statistic is

$$\chi^2 = \frac{n\left(|ad - bc| - \frac{n}{2}\right)^2}{(a+c)(b+d)(a+b)(c+d)}$$

$$\chi^2 = \frac{20\left(|1 \times 6 - 6 \times 7| - \frac{20}{2}\right)^2}{(1+7)(6+6)(1+6)(7+6)}$$

$$\chi^2 = \frac{20(36-10)^2}{8 \times 12 \times 7 \times 13} = \frac{13520}{8736} = 1.55$$

Number of degrees of freedom $= (r-1)(c-1) = (2-1)(2-1) = 1$

χ^2 table value for 1 d.f at $\alpha = 0.05$ is 3.841

$\Rightarrow \chi^2_{cal} < \chi^2_{tab}$

Therefore, H_0 is accepted.

Hence, Proportion of profit and licensor types are independent .

12.9 Lets sum up

In this chapter we have learnt the following:

Chi-square Distribution and its properties.

Goodness of fit for Chi-square distribution.

Uses of the Chi-square test.

Pair of categorical variables can be summarized using contingency table..

The Chi-square test can compare an observed contingency table to an expected table and determine if the categorical variable are independent.

YATE'S Correction for Contingency table.

12.10 Unit End exercise

1. The following table is given

Eye colour in fathers	Eye colour in sons		Total
	Brown	Black	
Brown	230	148	378
Black	251	471	622
Total	381	619	1000

Test whether the colour of the son's eyes is associated with that of the fathers.

2. Fit a Poisson distribution to the following data and test the goodness of fit.

x	0	1	2	3	4	5	6
f	275	72	30	7	5	2	1

3. A large city fire department calculates that for any given instance, during any given 8 hrs shift, there is a 30% chance of recurring at least one fire alarm. Here is a random sampling of 60 days:

No. of shift during which alarms were received	0	1	2	3
No. of days	16	27	11	6

At 5% level of significance verify that binomial distribution fits the data.

4. Test the0 hypothesis that the following observations follow a Poisson distribution with mean 4. Use 5% as level of significance.

No.of call per hrs.	0	1	2	3	4	5
No. of hrs.	20	57	98	85	78	62

5. Genetic theory states that children having one parent of blood type A and other blood type B will always be one of three types A, AB, B and that proportions of these types will on average be 1:2:1. A report states that out of 300 children having one A parent and one B parent, 30% were found to be type A , 45% of type AB and remaining of type B. Test the hypothesis by Chi-square test.

6. Fit a Binomial distribution to the data:

x	0	1	2	3	4	5
f	38	144	342	287	164	25

And test for goodness of fit at the level of significance 0.05.

7. The table shows the relation between the performance in mathematics and IT, using 0.01 significance level.

Mathematics Marks	IT Marks			Total
	High	Medium	Low	
High	56	71	12	139
Medium	47	163	38	248
Low	14	42	85	141
Total	117	276	135	528

8. In an experiment on immunization of human from COVID-19 the following results were obtained.

	Died	Unaffected
Inoculated	12	26
Not inoculated	16	6

Examine the effect of vaccine in controlling susceptibility to COVID-19.

9. A die is thrown 120 times with the following results:

Face	1	2	3	4	5	6
Frequency	16	30	22	18	14	20

Test the hypothesis that the die is unbiased at level of 5% significance.

10. In a survey of 200 boys, of which 75 were intelligent, 40 had skilled fathers; while 85 of the unintelligent boys had unskilled fathers. Do these figures support the hypothesis that skilled fathers have intelligent boys? Use Chi-square test at 5% level of signification.

11. Among 64 offspring's of a certain cross between guinea pigs, 34 were red, 10 were black and 20 were white. According to the genetic model, these numbers should be in the ratio 9:3:4, are the data consistent with the model at the 5% level?

12. Opinion about promotions, to be dependent on published work by persons interested in teaching or research was taken and displayed as below

Interest	Promotion dependent On published work		Total
	Agree	Disagree	
Teaching	90	10	100
Research	70	30	100
Total	160	40	200

Examine whether the promotion dependent on published work and interest.

12.11 Reference

Fundamentals of mathematical statistics by S.C. Gupta and V.K. Kapoor

A Guide to Chi-Squared Testing Priscilla E. Greenwood, Michael S. Nikulin

❖❖❖

13

LINEAR PROGRAMMING

Unit Structure

13.0 Objectives

This chapter would make you understand the following concepts:

- Sketching the graph for linear equations.
- Formulate the LPP
- Conceptualize the feasible region.
- Solve the LPP with two variables using graphical method.

13.1 Introduction

Linear programming (LP, also called linear optimization) is a method to achieve the best outcome (such as maximum profit or lowest cost) in a mathematical model whose requirements are represented by linear relationships. Linear programming is a special case of mathematical programming (also known as mathematical optimization).

13.2 Common terminology for LPP

Linear programming is a mathematical concept used to determine the solution to a linear problem. Typically, the goal of linear programming is to maximize or

minimize specified objectives, such as profit or cost. This process is known as optimization. It relies upon three different concepts: variables, objectives, and constraints.

"Linear programming is the analysis of problems in which a linear function of a number of variables is to be maximized (or minimized) when those variables are subject to a number of restrains in the form of linear inequalities."

Objective Function : The linear function, which is to be optimized is called objective function The objective function in linear programming problems is the real-valued function whose value is to be either minimized or maximized subject to the constraints defined on the given LPP over the set of feasible solutions. The objective function of a LPP is a linear function of the form $Z = a_1x_1 + a_2x_2 + a_3x_3 \ldots . a_nx_n$

Decision Variables : The variables involved in LPP are called decision variables denoted them as (x, y)or (x_1, x_2) etc. here its refer to some quantity like units, item production on sold quantity, Time etc.

Constraints : The constraints are limitations or restrictions on decision variables. they are expressed in linear equalities or inequalities i.e. $=, \leq, \geq$

Non-negative Constraints : This is a condition in LPP specifies that the value of variable being considered in the linear programming problem will never be negative. It will be either zero or greater than zero but can never be less than zero, Thus it is expressed in the form of $x \geq 0$ and $y \geq 0$.

Feasible Solution : A feasible solution is a set of values for the decision variables that satisfies all of the constraints in an optimization problem. The set of all feasible solutions defines the feasible region of the problem. in graph the overlapping region is called feasible region

Optimum solution : An optimal solution to a linear program is the solution which satisfies all constraints with maximum or minimum objective function value.

13.3 Mathematical Formulation of L.P.P

To write mathematical formulation of L.P.P following steps to be remembered.

Step 1 : Identify the variables involved in LPP (i.e. Decision variables) and denote them as (x, y)or (x_1, x_2).

Step 2 : Identify the objective function and write it as a mathematically in terms of decision variables.

Step 3 : Identify the different constraints or restrictions and express them mathematically

Example 13.3.1 :

A bakery produces two type of cakes I and II using raw materials R_1 and R_2. One cake of type I is produced by using 4 units of raw material R_1 and 6 units of raw material R_2 and one cake of type II is produced by using 5 units of raw material R_1 and 9 units of raw material R_2. There are 320 units of R_1 and 540 units R_2 in the stock. The profit per cake of type I and type II is Rs. 200 and Rs. 250 respectively. How many cakes of type I and type II be produced so as to maximize the profit? formulate the L.P.P

Solution:

Let x be the number of cakes of type I and y be the number of cakes of type II to be produce to get maximum profit.

since the production value is never negative

$\therefore \boldsymbol{x \geq 0, y \geq 0}$

This is non- negative constraints.

$\therefore$ The profit earned by selling 1 cake of type I is Rs. 200. Hence the profit earned by selling x cakes is Rs. $200x$.
Similarly, the profit earned by selling 1 cake of type II is Rs. 250 and hence profit earned by selling y cake is Rs. $250y$.

$\therefore$ The Profit earned is

$$\boldsymbol{Z = 200x + 250y}$$

This is objective function.

Now, after reading the given data carefully we can construct the following table

Cake Type	**Raw Material required**		**Profit**
	R_1 **(units)/Cake**	R_2 **(Units)/Cake**	
I	4	6	**200**
II	5	9	**250**
Availability (units)	**320**	**540**	**-**

∴ According to the table

∴ 1 Cake of type I consumes 4 units of R_1 hence x cakes of type I will consume $4x$ units of R_1 and one cake of type II consume 5 units of R_1 hence y cakes of type II will consume $5y$ units of R_1. But maximum number of units available of R_1 is 320. Hence, the constraint is

$4x + 5y \leq 320.$

Similarly, 1 Cake of type I consumes 6 units of R_2 hence x cakes of type I will consume $6x$ units of R_2 and one cake of type II consume 9 units of R_2 hence y cakes of type II will consume $9y$ units of R_2. But maximum number of units available of R_2 is 540. Hence, the constraint is

$6x + 9y \leq 540$

Hence the mathematical formulation of the given L.P.P is to

Maximize $Z = 200x + 250y$

Subject to,

$$\mathbf{4x + 5y \leq 320}$$

$$\mathbf{6x + 9y \leq 540}$$

$$\mathbf{x \geq 0, y \geq 0}$$

Example 13.3.2 :

A manufacture produce Ball pen and Ink pen each of which must be processed through two machines A and B. Machine A has maximum 220 hours available and machine B has maximum of 280 hours available . Manufacturing a Ink pen requires 6 hours on machine A and 3 hours on machine B. Manufacturing a Ball pen requires 4 hours on machine A and 10 hours on machine B. If the profit are Rs. 55 for Ink pen and Rs. 75 for Ball pen. Formulate the LPP to have maximum profit.

Solution :

Let Rs. Z be the profit, Which can be made by manufacturing and selling say 'x' number of Ink pens and 'y' number of Ball pens.

Here variable x and y are decision variables.

Since profit per Ink pen and ball pen is Rs. 55 and Rs. 75 respectively and we want to maximize the Z, Hence the objective function is

Max $Z = 55x + 75y$

We have to find x and y that maximize Z

We can construct the following tabulation form of given data:

Machine	Time in hours required for		Maximum available time in hours
	1 Ink pen	1 Ball pen	
A	6	4	220
B	3	10	280

A Ink pen requires 6 hr on machine A and Ball pen requires 4 hr on machine A and maximum available time of machine A is 220 hr.

1st Constraint is

$6x + 4y \leq 220$

Similarly, A Ink pen requires 3 hr on machine B and Ball pen requires 10 hr on machine B and maximum available time of machine B is 280 hr.

2nd Constraint is

$3x + 10y \leq 280$

Here the production of Ball pen and Ink pen can not be negative:

we have Non-Negative constraints as $x \geq 0, y \geq 0$

Hence the required formulation of LPP is as follows:

Max $Z = 55x + 75y$

Subject to,

$\mathbf{6x + 4y \leq 220}$

$\mathbf{3x + 10y \leq 280}$

$\mathbf{x \geq 0, y \geq 0}$

Example 13.3.3 :

In a workshop 2 models of agriculture tools are manufactured A_1 and A_2. Each A_1 requires 6 hours for I processing and 3 hours for II processing. Model A_2 requires 2 hours for I processing and 4 hours for II processing. The workshop has 2 first processing machines and 4 second processing machine each machine of I processing units works for 50 hrs a week. Each machine in II processing units

works for 40 hrs a week. The workshop gets Rs. 10/- profit on A_1 and Rs. 14/- on A_2 on sale of each tool. Determine the maximum profit that the work shop get by allocating production capacity on production of two types of tool A and B

Solution:

Decision Variables :

1. Let the number of units of type A_1 model tools be x.

2. Let the number of units of type A_2 model tools be y.

Objective function :

The objective of the workshop is to obtain maximum profit by allocating his production capacity between $A_1$and A_2 and 10 Rs. per unit profit on model A_1 and 14 Rs. on model A_2

$\therefore Z = 10x + 14y$

Constraints :

1. For processing of A_1 tool and A_2 tools require $6 + 2 = 8$ hrs in I Processing unit $= 6x + 2y$
2. For processing of A_1type of tools and A_2 type requires $3 + 4 = 7$ hrs in II processing unit $= 3x + 4y$

Total machine hours available in I processing unit = $2 \times 50 = 100$ hrs per week.

Total machine hours available in II processing unit = $4 \times 40 = 160$ hrs per week.

Considering the time constraint, the constrain function can be written in the following way:

$6x + 2y \leq 100$

$3x + 4y \leq 160$

Non-negative constraint :

There is no possibility of negative production in the workshop

$\therefore$ The non-negative function will be

$x \geq 0, y \geq 0$

Mathematical form of the production of 2 types of tools in the work shop to maximize profits under given constraints will be in the following way.

Maximize $Z = 10x + 14y$

Subject to,

$$6x + 2y \leq 100$$

$$3x + 4y \leq 160, x \geq 0, y \geq 0$$

Example 13.3.4 :

Diet for a sick person must contain at least 400 units of vitamins, 500 units of minerals and 300 calories. Two foods F_1 and F_2 cost Rs. 2 and Rs. 4 per unit respectively. Each unit of food F_1 contains 10 units of vitamins, 20 unit of minerals and 15 calories, whereas each unit of food F_2 contains 25 units of vitamins, 10 units of minerals and 20 calories. Formulate the L.P.P. to satisfy sick person's requirement at minimum cost.

Solution :

After reading this carefully, Tabulation form of give data:

Micronutrients	F_1	F_2	**Minimum units requirement**
vitamins	10	25	400
minerals	20	10	500
calories	15	20	300
Cost	2	4	

Decision Variables :

Let x for F_1 and y for F_2

Objective function :

We have to find minimum the cost for a diet hence the objective function in terms of decision variables is

Minimize $Z = 2x + 4y$

Constraints :

First constraints : $10x + 25y \geq 400$

Second Constraints : $20x + 10y \geq 500$

Third constraints : $15x + 20y \geq 300$

Non-negative constraint

There is no possibility of negative food quantity of the diet

hence, $x \geq 0, y \geq 0$

Mathematical formulation of LPP can be written as

Minimize $Z = 2x + 4y$

Subject to,

$$10x + 25y \geq 400$$

$$20x + 10y \geq 500$$

$$15x + 20y \geq 300, \quad x \geq 0, y \geq 0$$

Example 13.3.5 :

A garden shop wishes to prepare a supply of special fertilizer at a minimal cost by mixing two fertilizer, A and B. The mixture is contains : at least 45 units of phosphate, at least 36 units of nitrate at least 40 units of ammonium. Fertilizer A cost the shop Rs 0.97 per Kg. fertilizer B cost the shop Rs.1.89 per Kg. Fertilizer A contains 5 units of phosphate and 2 units of nitrate and 2 units of ammonium, fertilizer B contains 3 units of phosphate and 3 units of nitrate and 5 units of ammonium. How many pounds of each fertilizer should the shop use in order to minimum their cost ?

Solution:

After reading this carefully, Tabulation form of give data:

Contains	**Fertilizer type**		**Minimum units requirement**
	A	**B**	
Phosphate	5	3	45
Nitrate	2	3	36
Ammonium	2	5	40
Cost	0.97	1.89	–

Decision Variables :

Let x for A and y for B

Objective function :

We have to find minimum the cost, Hence the objective function in terms of decision variables is

Minimize $Z = 0.97x + 1.89y$

Constraints :

First constraints : $5x + 3y \geq 45$

Second Constraints : $2x + 3y \geq 36$

Third constraints : $2x + 5y \geq 40$

Non-negative constraint

There is no possibility of negative supply of fertilizer

hence, $x \geq 0, y \geq 0$

Mathematical formulation of LPP can be written as

Minimize $Z = 0.97x + 1.89y$

Subject to,

$\mathbf{5x + 3y \geq 45}$

$\mathbf{2x + 3y \geq 36}$

$\mathbf{2x + 5y \geq 40,}$

$\mathbf{x \geq 0, y \geq 0}$

Example 13.3.6 :

A printing company prints two types of magazines A and B the company earns Rs. 25 and Rs. 35 on each copy of magazines A and B respectively. The magazines are processed on three machines. Magazine A requires 2 hours on machine I, 4 hours on machine II and 2 hours on machine III. Magazine B requires 3 hours on machine I, 5 hours on machine II and 3 hours on machine III. Machines I, II and III are available for 35, 50 and 70 hours per week respectively formulate the L.P.P. so as to maximize the total profit of the company.

Solution :

Tabulation form of give data:

Machine	**Magazine**		**Maximum availability**
	A	**B**	
I	2	3	35
II	4	5	50
III	2	3	70

Decision Variables :

Let x number of copies of magazine A and y number of copies of magazine B to be printed to get maximum profit.

Objective function :

We have to Maximize the Profit, Hence the objective function in terms of decision variables is

$$\text{Maximize } Z = 25x + 35y$$

Constraints :

First constraints : $2x + 3y \leq 35$

Second Constraints : $4x + 5y \leq 50$

Third constraints : $2x + 3y \leq 70$

Non-negative constraint

There is no possibility of negative production of magazines

∴ The non-negative function will be

$x \geq 0, y \geq 0$

Mathematical formulation of LPP can be written as

Maximize $Z = 25x + 35y$

Subject to,

$\mathbf{2x + 3y \leq 35}$

$\mathbf{4x + 5y \leq 50}$

$\mathbf{2x + 3y \leq 70}$

$\mathbf{x \geq 0, y \geq 0}$

Example 13.3.7 :

A food processing and distributing units has 3 production units A,B,C in three different parts of a city. They have five retails out lest in the city P, Q, R, S and T to which the food products are transported regularly. Total stock available at the production units is 500 units which is in the following ways : A=200 units; B=120 units and C= 180 units. Requirement at the retails outlets of the industry are: A=125, B=150, C=100, D=50, E=75. Cost of transportation of products from different production centers to different retail outlets is in the following way:

	P	Q	R	S	T
A	2	12	8	5	6
B	6	10	10	2	5
C	12	18	20	8	9

How the industry can minimize the cost on transportation of products. Formulate the linear programming problem.

Solution :

The objective of the industry is to minimize the possible cost on transportation

Let Z be the objective function.

Let x_1, x_2, x_3, x_4 and x_5 are decision variables

Tabulation form of the given data

Production centers	**Decision variables**	**Retails out lets**					**No. of units can be supplied**
		x_1	x_2	x_3	x_4	x_5	
A	x_1	2	12	8	5	6	200
B	x_2	6	10	10	2	5	120
C	x_3	12	18	20	8	9	180
Units of demand		**125**	**150**	**100**	**50**	**75**	**500**

Objective function :

The objective is to minimize the cost

$$\text{Minimize cost} = 2x_{11} + 12x_{12} + 8x_{13} + 5x_{14} + 6x_{15} + 6x_{21} + 10x_{22} + 10x_{23} + 2x_{24} + 5x_{25} + 12x_{31} + 18x_{32} + 20x_{33} + 8x_{34} + 9x_{35}$$

Constraint Function :

Supply constraint,

In the problem requirement at the retail out lets and the supply at the production units is the same. therefore supply constraint will be

$$A = x_{11} + x_{12} + x_{13} + x_{14} + x_{15} = 200$$

$$B = x_{21} + x_{22} + x_{23} + x_{24} + x_{25} = 120$$

$$C = x_{31} + x_{32} + x_{33} + x_{34} + x_{35} = 180$$

Demand Constraints :

P $= x_{11} + x_{12} + x_{13} = 125$

Q $= x_{21} + x_{22} + x_{23} = 150$

R $= x_{31} + x_{32} + x_{33} = 100$

S $= x_{41} + x_{42} + x_{43} = 50$

T $= x_{51} + x_{52} + x_{53} = 75$

Non- Negative function:

$x_{11}, x_{12}, x_{13}, x_{14}, x_{15} \geq 0$

13.4 Graphical Method

Graphical method, the inequality constraints are taken as equalities. Each equality constraints is drawn on the graph paper which forms a straight line. Lines are drawn equal to the constrains. Then the region which satisfies all inequality is located. this region is known as feasible region. Solution determine with regard to this region is called the feasible solution. Accordingly to Hadley "if an optimum (maximum or minimum) value of a linear programming problem exits then it must correspond to one of the corner points of the feasible region" Feasible region corresponding to a linear programming problem can be located by constructing graph as given below.

Steps for solving L.P.P. graphically

1. Formulate the mathematical linear programming problem. there will be 2 variables x and y.

2. Since both x and y are non negative, graphic solution will be restricted to the first quadrant.

3. Choose an appropriate scale for x and y axis.

4. Each inequality in the constraints equation can be written as equality. Example. $x + y \leq 70$ them make it $x + y = 70$

5. Given any arbitrary value to one variable and get the value of other variable by solving the equation. Similarly given another arbitrary value to the variable and find the corresponding value of the other variable. Example $x + y = 7$ at any point on x axis y will be 0 the $x + y = 7$ and $x = 7$ then at any point on y axis x will be zero $y = 5$, thus $(0,5)$ is a point on x axis and $(0,5)$ is a point on y axis that satisfies the equation.

6. Now plot these two sets of values connect these points by straight line. That divides the first quadrant into two parts. Since the constraints is an inequality, one of the two sides satisfies inequality

7. Repeat these steps for every constraints stated in the linear programming problem.

8. There forms a common area called feasible area.

9. For greater than or greater than equal to constraints, the feasible region will be the area which lies above the constrains.

13. For less than or less than equal, the area is below these lines.

Example 13.4.1 :

Solve the following linear programming by graphical method.

Maximize $Z = 5x + 6y$

Subject to, $2x + 4y \leq 16$

$3x + y \leq 12$

$3x + 3y \leq 24,\ x \geq 0, y \geq 0$

Solution :

By converting the inequality equation to equality equation we get:

$2x + 4y = 16$ ...(equation 1)

$3x + y = 12$ ...(equation 2)

$3x + 3y = 24$ ...(equation 3)

Equation (1)

$2x + 4y = 16$, When $x = 0$, then $y = \frac{16}{4} = 4$, When $y = 0$ then $x = \frac{16}{2} = 8$,

x	0	8
y	4	0

Equation (2)

$3x + y = 12$,When $x = 0$ then $y = 12$, When $y = 0$ then $x = \frac{12}{3} = 4$,

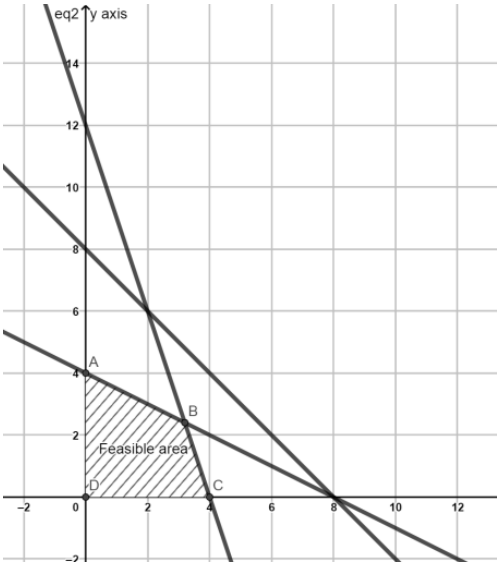
x	0	4
y	12	0

Equation (3)

$3x + 3y = 24$,When $x = 0$ then $y = \frac{24}{3} = 8$, When $y = 0$ then $x = \frac{24}{3} = 8$,

x	0	8
y	8	0

By above equation coordinates we can draw the straight lines to obtain feasible area



Thus the feasible region is OABC

Where $A(0,4)$, $B(3.2,2.4)$, and $C(4,0)$

Consider, $Z = 5x + 6y$

at $A(0,4)$, $Z = 5(0) + 6(4) = 24$

at $B(3.2,2.4)$, $Z = 5(3.2) + 6(2.4) = 30.4$

at $C(4,0)$, $Z = 5(4) + 6(0) = 20$

Thus Z is maximum at point B, Thus the solution is $x = 3.2$ and $y = 2.4$

Example 13.4.2 :

Solve LPP graphically, Maximum $Z = 10x + 5y$

Subject to, $x + y \leq 5$,

$2x + y \leq 6$,

$x \geq 0, y \geq 0$

Solution : Converting the given constraints in to equations (or ignore the inequality to find the co-ordinates for sketching the graph) we get,

$x + y = 5$ --- Equation (1)

$2x + y = 6$ --- Equation (2)

Consider the equation (1)

$x + y = 5$, When $x = 0$ then $y = 5$ and $y = 0$ then $x = 5$

x	0	5
y	5	0

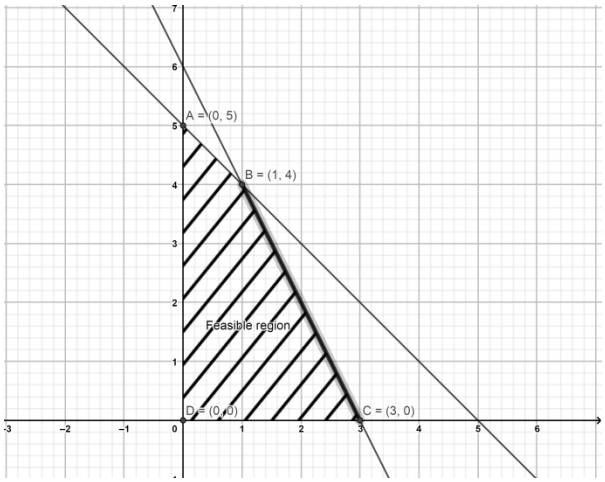
Plot $(0,5)$ and $(5,0)$ on the graph. Join the points by a straight line and shade the region represented by, $x + y \leq 5$

Then, Consider the equation (2)

$2x + y = 6$, When $x = 0$ then $y = 6$ and $y = 0$ then $x = 3$

x	0	3
y	6	0

Plot $(0,6)$ and $(3,0)$ on the graph. Join the points by a straight line and shade the region represented by, $2x + y \leq 6$



From the graph we get the feasible region ABCD is a feasible region where $D(0,0), A(0,5), B(1,4), C(3,0)$.

Consider the value of Z at different values of corner points of feasible region.

Corner point	$Z = 10x + 5y$
D	0
A	25
B	30
C	30

∴ Here Z attained maximum value at two points which is B and C

in this case any point which lies on set BC is also another optimal solution

∴ There are infinitely many optimal solution.

Example 13.4.3 :

Solve the following LPP graphically

Min $Z = 2x + 3y$

Subject to, $x + y \leq 3$

$2x + 2y \geq 10$

$x \geq 0, y \geq 0$

Solution :

Converting the given constraints in to equations (or ignore the inequality to find the co-ordinates for sketching the graph) we get,

$x + y = 3$ --- Equation (1)

$2x + 2y = 10$ --- Equation (2)

Consider the equation (1) $x + y = 3$, When $x = 0$ then $y = 3$ and $y = 0$ then $x = 3$

x	0	3
y	3	0

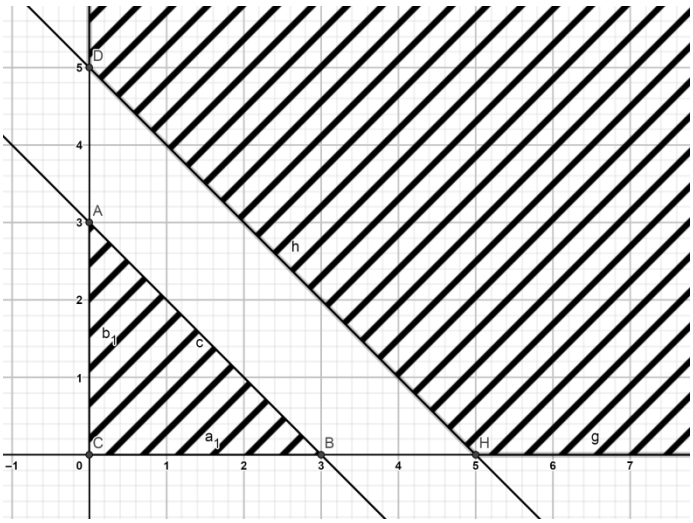
Plot (0,3) and (3,0) on the graph. Join the points by a straight line and shade the region represented by, $x + y \leq 3$

Then, Consider the equation (2)

$2x + 2y = 10$, When $x = 0$ then $y = 5$ and $y = 0$ then $x = 5$

x	0	5
y	5	0

Plot $(0,5)$ and $(5,0)$ on the graph. Join the points by a straight line and shade the region represented by, $2x + 2y \geq 10$



∴ This LPP does not have common feasible region and hence no optimum solution.

Example 13.4.4 :

Consider a calculator company which produce a scientific calculator and graphing calculator. long-term projection indicate an expected demand of at least 1000 scientific and 800 graphing calculators each month. Because of limitation on production capacity, no more than 2000 scientific and 1700 graphing calculators can be made monthly. To satisfy a supplying contract a total of at least 2000 calculator must be supplied each month. if each scientific calculator sold result in Rs. 120 profit and each graphing calculator sold produce Rs. 150 profit, how many of each type of calculator should be made monthly to maximize the net profit ?

Solution :

Decision variables

Let x is the number of scientific calculators produced and y is the number of graphing calculators produced.

Objective function

We have to find maximum profit. Hence objective function in a terms of decision variable is,

Maximize $Z = 120x + 150y$

Constraints

at least 1000 scientific calculators : $x \geq 1000$

at least 800 graphing calculators : $y \geq 800$

no more than 2000 scientific calculators : $x \leq 2000$

no more than 1700 graphing calculators : $y \leq 1700$

a total of at least 2000 calculators : $x + y \geq 2000$

and non-negative constraints : $x \geq 0, y \geq 0$

Hence the formulation of LPP can be written as

Maximize $Z = 120x + 150y$

Subject to, $1000 \leq x \leq 2000$

$800 \leq y \leq 1700$

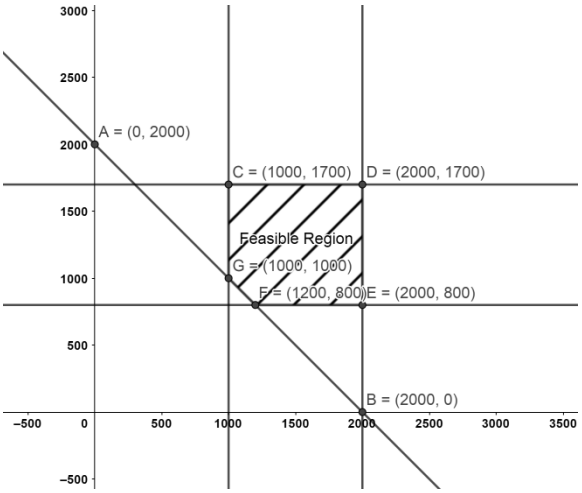
$x + y \geq 2000$

$x \geq 0, y \geq 0$

Consider, $x + y \geq 2000$

if $x = 0$ then $y = 2000$ and when $y = 0$ then $x = 2000$

∴ The point are $A(0,2000)$ and $B(2000,0)$



From the graph CDEFG is feasible region

To find the value of point G, Consider $x = 1000$ and $x + y = 2000$,

$\Rightarrow y = 1000$, Point G is (1000,1000)

To find the value of point F, Consider $y = 800$ in $x + y = 2000$,

$\Rightarrow y = 1200$, Point G is (1200,800)

Here CDEFG is feasible region.

Where $C(1000,1700), D(2000,1700), E(2000,800), F(1200,800)$ and $G(1000,1000)$

Now, $Z = 120x + 150y$

at $C(1000,1700)$, $Z = 375000$

at $D(2000,1700)$, $Z = 495000$

at $E(2000,800)$, $Z = 360000$

at $F(1200,800)$, $Z = 264000$

at $G(1000,1000)$, $Z = 270000$

$\therefore Z$ is maximum at point $D(2000,1700)$

$\therefore$ The maximum value of $120x + 150y$ is 495000 at $(2000,1700)$.

$\therefore$ 2000 scientific and 1700 graphing calculators should be made monthly to maximize the net profit.

Example 13.4.5 :

Solve the following LPP graphically,

Minimize $Z = 25x + 10y$

Subject to, $10x + 2y \geq 20$

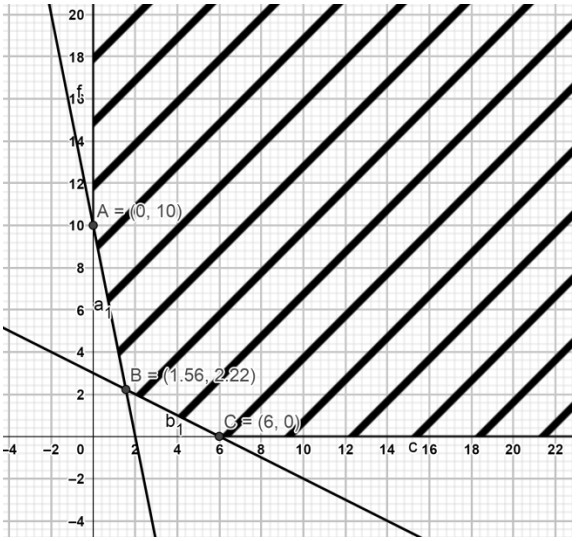
$x + 2y \geq 6, \quad x \geq 0, y \geq 0$

Solution :

Consider the equation $10x + 2y = 20$, when $x = 0$ then $y = 10$ and when $y = 0$ then $x = 2$.

Consider the for 2nd equation $x + 2y = 6$, when $x = 0$ then $y = 3$ and when $y = 0$ then $x = 6$.

Plotting the graph



From the graph we get the feasible region where $A(0,10), B(1.56,2.22), C(6,0)$,.

For Minimize $Z = 25x + 10y$

point	$Z = 25x + 10y$
A	100
B	61.2
C	150

∴ Minimum $Z = 61.2$ at point B i.e. $x = 1.56$ and $y = 2.22$

13.5 Summary

- The objective of this course is to present Linear Programming. Activities allow to train modeling practical problems in Linear programming. Various tests help understanding how to solve a linear program
- Linear programming is a mathematical technique for solving constrained maximization and minimization problems when there are many constraints and the objective function to be optimized as well as the constraints faced are linear (i.e., can be represented by straight lines). Linear programming has been applied to a wide variety of constrained optimization problems. Some of these are: the selection of the optimal production process to use to produce a product, the optimal product mix to produce, the least-cost input combination to satisfy some minimum product requirement, the marginal contribution to profits of the various inputs, and many others.

13.6 References

Following books are recommended for further reading :-

- Numerical Methods for Engineers - *Steven C. Chapra, Raymond P. Canale*; Tata Mc Graw Hill.
- Quantitative Techniques - *Dr. C. Satyadevi* ; S. Chand.

13.7 Exercise

Q. (1) A firm manufactures headache pills in two size A and B. size A contains 3 grains of aspirin, 6 grains of bicarbonate and 2 grain of codeine. Size B contains 2 grains of aspirin, 8 grains of bicarbonate and 10 grain of codeine. It is found by users that it requires at least 15 grains of aspirin, 86 grains of bicarbonate and 28 grain of codeine of providing immediate effect. it is required to determine the least number of pills a patient should take to get immediate relief. Formulate the problem as a standard LPP.

Q. (2) An animal feed company must produce 200lbs of mixture containing the ingredients A and B. A cost Rs. 5 per Lb. And B cost Rs. 10 per lb. Not more than 100 lbs. of A can used and minimum quantity to be used for B is 80lbs. Find how much of each ingredient should be used if the company wants to minimize the cost. Formulate the problem as a standard LPP.

Q. (3) A painter make two painting A and B. He spends 1 hour for drawing and 3 hours for coloring the painting A and he spends 3 hours for drawing and 1 hour for coloring the painting B. he can spend at most 8 hours and 9 hours for drawing and coloring respectively. The profit per painting of Type A is Rs. 4000 and that of type B is Rs. 5000. formulate as LPP to maximize the profit

Q. (4) A gardener wanted to prepare a pesticide using two solutions A and B. the cost of 1 liter of solution A is Rs. 2 and the cost of 1 liter of solutions B is Rs. 3. He wanted to prepare at least 20 liter of pesticide. The quality of solution A available in a shop is 12 liter and solution B is 15 liter. How many liter of pesticide the gardener should prepare so as to minimize the cost ? formulate the LPP

Q. (5) A bakery produce two kinds of biscuits A and B, using same ingredients L1 and L2. The ingredients L1 and L2 in biscuits of type A are in the ratio 4:1 and in biscuits of type B are in the ratio 9:1. The Profit per Kg for biscuits of type A and B is Rs. 8 per Kg and Rs. 10 per Kg respectively. The bakery had 90 Kg of L1 and 20 Kg of L2 in stock. How many Kg of biscuits A and B be produced to maximize total profit ?

Q. (6) Solve graphically the following LPP

Maximize $Z = 2x + 3y$

subject to, $x + 2y \geq 6$

$2x - 5y \leq 1, x \geq 0, y \geq 0$

Q. (7) Solve graphically the following LPP

Maximize $Z = 2x + 5y$

subject to, $4x + 2y \leq 80$

$2x + 5y \leq 180, x \geq 0, y \geq 0$

Q. (8) Solve graphically the following LPP

Maximize $Z = 2x + y$

subject to, $4x - y \leq 3$

$2x + 5y \leq 7, x \geq 0, y \geq 0$

Q. (9) Solve graphically the following LPP

Maximize $Z = 400x + 500y$

subject to, $x + 3y \leq 8$

$3x + y \leq 9, x \geq 0, y \geq 0$

Q. (10) Solve graphically the following LPP

Maximize $Z = 5000x + 4000y$

subject to, $6x + 4y \leq 24$

$x + 2y \leq 6$

$-x + y \leq 1$

$y \leq 2,\ x \geq 0, y \geq 0$

Q. (11) Solve graphically the following LPP

Minimize $Z = 30x + 50y$

subject to, $3x + 4y \geq 300$

$x + 3y \geq 210, x \geq 0, y \geq 0$

❖❖❖