ENGINEERING MATHEMATICS

ENGINEERING MATHEMATICS
First Edition

Amit K Awasthi
Associate Professor, PSIT Kanpur

Sanjay Chaudhary
Professor, PSIT Kanpur

darbose

Darbose Prentice Hall, India

Published by Darbose, Inc., India.

Library of Congress Cataloging-in-Publication Data:

A Textbook of Engineering Mathematics/Darbose

1. Advance Mathematics. 2. Engineering Mathematics

Engineering Mathematics

UPTU Textbook Series.

Printed and distributed by Bhakti Prakashan.
Printed in the India.

10 9 8 7 6 5 4 3 2 1

To my mother
She lost her fight with
cancer on 23 Jun 2009
during this project was
going on.

Amit K Awasthi

CONTENTS

PREFACE

Mathematics is a necessary avenue to scientific knowledge which opens new vistas of mental activity. A sound knowledge of Engineering Mathematics is must for the modern engineer to attain new dimensions in all aspects of engineering practices. Applied mathematics is alive and very vigorous. That ought to be reflected in our teaching. In our own teaching we became convinced that the textbook is crucial. It must provide a framework into which the applications will fit. A good course has a clear purpose, and you can sense that it is there. It is a pleasure to teach a subject when it is moving forward, and this one is-but the book has to share in that spirit and help to establish it.

This book is a self-contained, comprehensive volume covering the entire ambit of the course of Engineering Mathematics for III Semester B.Tech program of UPTU. This text covers following contents

Unit-I: Function of Complex variable. Analytic function, C-R equations, Cauchys integral theorem, Cauchys integral formula for derivatives of analytic function, Taylors and Laurents series, singularities, Residue theorem, Evaluation of real integrals of the type $\int_0^{2\pi} f(\cos\theta, \sin\theta)d\theta$ and $\int_{-\infty}^{\infty} f(x)dx$.

Unit–II: Statistical Techniques - I. Moments, Moment generating functions, Skewness, Kurtosis, Curve fitting, Method of least squares, Fitting of straight lines, Polynomials, Exponential curves etc., Correlation, Linear, non linear and multiple regression analysis, Probability theory.

Unit–III: Statistical Techniques - II. Binomial, Poisson and Normal distributions, Sampling theory (small and large), Tests of significations: Chi-square test, t-test, Analysis of variance (one way) , Application to engineering, medicine, agriculture etc. Time series and forecasting (moving and semi-averages), Statistical quality control methods, Control charts,$\overline{X}$, R, p, np, and c charts.

Unit–IV: Numerical Techniques I. Zeroes of transcendental and polynomial equation using Bisection method, Regula-falsi method and Newton-Raphson method, Rate of convergence of above methods. Interpolation: Finite differences, difference tables, Newtons forward and backward interpolation , Lagranges and Newtons divided difference formula for unequal intervals.

Unit–V: Numerical Techniques II. Solution of system of linear equations, Gauss- Seidal method, Crout method. Numerical differentiation, Numerical integration , Trapezoidal , Simpsons one third and three-eight rules, Solution of ordinary differential (first order, second order and simultaneous) equations by Eulers, Picards and forth-order Runge- Kutta mehthods.

This book is written in a lucid, easy to understand language. Each topic has been thoroughly covered in scope, content and also from the examination point of view. For each topic, several worked out examples, carefully selected to cover all aspects of the topic, are presented. This is followed by practice exercise with answers to all the problems and hints to the difficult ones. Students may visit the book website http://www.darbose.com/em/ for further discussion and updates.

We are hopeful that this exhaustive work will be useful to both students as well as teachers. If you have any queries, please feel free to write at: awasthi@psit.in

In spite of our best efforts, some errors might have crept in to the book. Report of any such error and all suggestions for improving the future editions of the book are welcome and will be gratefully acknowledged.

AUTHORS

Kanpur, UP, India
August, 2009

ACKNOWLEDGMENTS

Partial Contribution

Dr Atul Chaturvedi
Dr Shyam Sundar
Mr Puneet Dubey
Ms Keerti Srivastava
Ms Divya Shukla
Dr Rajan K Sharma

We are thankful to all partial contributors for their timely comments as well as thorough review of the sections. To all these wonderful people I owe a deep sense of gratitude especially now that this project has been completed.

Authors

UNIT I

FUNCTION OF COMPLEX VARIABLES

CHAPTER 1

ANALYTIC FUNCTIONS

The theory of functions of complex variable is utmost important in solving a large number of problems in the field of engineering. Many complicated integrals of real functions are solved with the help of functions of a complex variable.

1.1 INTRODUCTION

Let Z and W be two non-empty set of complex numbers. A rule f assigns to each element $z \in Z$, a unique $w \in W$, is called as complex function. i.e.,

$$f : Z \to W$$

We may also write,

$$w = f(z)$$

Here z and w are complex variables. As $z = x + iy$, x and y are independent real variables. Let $w = u + iv$, a function of z, which implies that u and v are function x and y as z is function x and y. i.e,

$$u \equiv u(x,y)$$

$$v \equiv v(x,y)$$

Thus

$$w = f(z) = u(x,y) + iv(x,y)$$

There may be some relations such that, to a value of z, there may correspond more than one value of w, such as $w = \sqrt{z}$ has two values of w for each $z \in Z$. We call such relations *multivalued functions*. On the other hand in the usual sense functions returning single value of w for each $z \in Z$ are called *single valued functions*. We use function for single valued functions throughout the text.

1.1.1 Limits

A function $f(z)$ is said to have limit A as $z \to a$, if $f(z)$ is defined in deleted neighborhood[1] of a and if $\forall \varepsilon, \exists \delta > 0$, such that $|f(z) - A| < \varepsilon$ whenever $0 < |z - a| < \delta$.

If $f(z) = f(x+iy) = u(x,y) + iv(x,y)$. Let $a = x_0 + iy_0$ Then,

$$\lim_{z \to a} f(z) = A = \alpha + i\beta$$

$$\Leftrightarrow$$

$$\lim_{x \to x_0, y \to y_0} u(x,y) = \alpha$$

and

$$\lim_{x \to x_0, y \to y_0} v(x,y) = \beta$$

1.1.2 Continuity

$f(z)$ is said to be continuous at $z = a$, if $f(a)$ is defined and $\lim_{z \to a} f(z) = f(a)$. In other words, Let function $f(z)$ of complex variable z is said to continuous at the point z_0, if for any given positive number ε, we can find a number δ such that $|f(z) - f(z_0)| < \varepsilon$, for all points z of the domain satisfying $|z - z_0| < \delta$.

Also, if $f(z) = f(x+iy) = u(x,y) + iv(x,y)$. Let $f(z)$ is continuous at $a = x_0 + iy_0$ Then u and v are separately continuous at the point $a = x_0 + iy_0$.

$f(z)$ is said to be continuous in domain if continuous at each point of that domain.

■ **EXAMPLE 1.1**

The function $z, Re(z), Im(z)$ and $|\overline{z}|$ are continuous in the entire plane.

1.1.3 Differentiability

The *complex derivative* is defined as,

$$\frac{d}{dz} f(z) = \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z}.$$

The complex derivative exists if this limit exists. This means that the value of the limit is independent of the manner in which $\Delta z \to 0$. If the complex derivative exists at a point, then we say that the function is *complex differentiable* there.

[1] A subset C_N of the complex plane containing z_0 is said to be neighborhood of z_0, if for some real number $\delta > 0$, the set $\{z \in C : |z - z_0| < \delta\} \subseteq C_N$. Further the set $C_N - \{z_0\}$ is called deleted neighborhood of z_0.

■ **EXAMPLE 1.2**

Show that $f(z) = \bar{z}$ is not differentiable.

Solution: Consider its derivative.

$$\frac{d}{dz} f(z) = \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z}.$$

$$\begin{aligned}\frac{d}{dz}\bar{z} &= \lim_{\Delta z \to 0} \frac{\overline{z + \Delta z} - \bar{z}}{\Delta z} \\ &= \lim_{\Delta z \to 0} \frac{\overline{\Delta z}}{\Delta z}\end{aligned}$$

First we take $\Delta z = \Delta x$ and evaluate the limit.

$$\lim_{\Delta x \to 0} \frac{\Delta x}{\Delta x} = 1$$

Then we take $\Delta z = i\Delta y$.

$$\lim_{\Delta y \to 0} \frac{-i\Delta y}{i\Delta y} = -1$$

Since the limit depends on the way that $\Delta z \to 0$, the function is nowhere differentiable. □

■ **EXAMPLE 1.3**

Prove that the function $f(z) = |z|^2$ is continuous everywhere but nowhere differentiable except at the origin.

Solution: Since $f(z) = |z|^2 = x^2 + y^2$, the continuity of the function $f(z)$ is evident because of the continuity of $x^2 + y^2$.

Let us consider its differentiability,

$$\begin{aligned}f'(z_0) &= \lim_{\delta z \to 0} \frac{f(z_0 + \delta z) - f(z_0)}{\delta z} \\ &= \lim_{\delta z \to 0} \frac{|z_0 + \delta z|^2 - |z_0|^2}{\delta z} \\ &= \lim_{\delta z \to 0} \frac{(z_0 + \delta z)\overline{(z_0 + \delta z)} - z_0\overline{z_0})}{\delta z} \\ &= \lim_{\delta z \to 0} \overline{z_0} + \delta\bar{z} + z_0\frac{\delta\bar{z}}{\delta z} \\ &= \lim_{\delta z \to 0} \overline{z_0} + z_0\frac{\delta\bar{z}}{\delta z} \qquad (\text{Since } \delta z \to 0 \Rightarrow \delta\bar{z} \to 0)\end{aligned}$$

Now at $z_0 = 0$, the above lim,it is clearly zero, so that $f'(0) = 0$. Let us now choose $z_0 \neq 0$, let

$$\begin{aligned}\delta z &= re^{i\theta} \\ \delta\bar{z} &= re^{-i\theta} \\ \Rightarrow \frac{\delta\bar{z}}{\delta z} &= e^{-2i\theta} = \cos 2\theta - i\sin 2\theta\end{aligned}$$

Above does not not tend to a unique limit as this limit depends upon θ. Therefore, the given function is not differentiable at any other non-zero value of z. □

EXAMPLE 1.4

If

$$f(z) = \begin{cases} \dfrac{x^3y(y-ix)}{x^6+y^2}, & z \neq 0 \\ = 0, & z = 0, \end{cases}$$

prove that $\dfrac{f(z)-f(0)}{z} \to 0$ as $z \to 0$ along any radius vector, but not $z \to 0$ in any manner.

Solution: Here, $y - ix = -i(x+iy) = -iz$. Now,

$$f'(z) = \lim_{z\to 0} \frac{f(z)-f(0)}{z} = \lim_{z\to 0} \frac{\frac{x^3y(y-ix)}{x^6+y^2} - 0}{z}$$
$$= \lim_{z\to 0} \frac{x^3yi}{x^6+y^2}$$

Now if $z \to 0$ along any radius vector, say $y = mx$, then

$$f'(z) = \lim_{z\to 0} \frac{f(z)-f(0)}{z} = \lim_{z\to 0} \frac{x^3(mx)i}{x^6+(mx)^2}$$
$$= \lim_{z\to 0} \frac{x^4im}{x^6+m^2x^2}$$
$$= \lim_{z\to 0} \frac{x^2im}{x^4+m^2} = 0 \quad \text{Hence.}$$

Now let us suppose that $z \to 0$ along the curve $y = x^3$, then

$$f'(z) = \lim_{z\to 0} \frac{f(z)-f(0)}{z} = \lim_{z\to 0} \frac{x^3(x^3)i}{x^6+(x^3)^2}$$
$$= \lim_{z\to 0} \frac{x^6 i}{x^6+x^6} = -\frac{i}{2}$$

along different paths, the value of $f'(z)$ is not unique (that is 0 along $y = mx$ and $-\dfrac{i}{2}$ along $y = x^3$). Therefore the function is not differentiable at $z = 0$. □

1.2 CAUCHY-RIEMANN EQUATIONS

Theorem 1. *The necessary and sufficient condition for the derivative of the function $f(z) = u + iv$, where u and v are real-valued functions of x and y, to exist for all values of z in domain D, are*

1. $\dfrac{\partial u}{\partial x} = \dfrac{\partial v}{\partial y}$ *and* $\dfrac{\partial u}{\partial y} = -\dfrac{\partial v}{\partial x}$.
2. $\dfrac{\partial u}{\partial x}, \dfrac{\partial u}{\partial y}, \dfrac{\partial v}{\partial x}, \dfrac{\partial v}{\partial y}$ *are continuous function of x, y in D.*

provided these four partial derivatives involved here should exist. The relations given in (1) is referred as .Cauchy and Riemann Equations (some times CR Equations).

Proof. **Necessary Condition** Let derivative of $f(z)$ exists, then

$$\begin{aligned} f'(z) &= \lim_{\delta z \to 0} \frac{f(z+\delta z) - f(z)}{\delta z} \\ &= \lim_{\delta x, \delta y \to 0,0} \frac{[u(x+\delta x, y+\delta y) + iv(x+\delta x, y+\delta y)] - [u(x,y) + iv(x,y)]}{\delta x + i\delta y} \end{aligned}$$

since $f'(z)$ exists, the limit of above equation should be finite as $(\delta x, \delta y) \to (0,0)$ in any manner that we may choose. To begin with, we assume that δz is wholly real, i.e. $\delta y = 0$ and $\delta z = \delta x$. This gives

$$\begin{aligned} f'(z) &= \lim_{\delta x \to 0} \frac{[u(x+\delta x, y) + iv(x+\delta x, y)] - [u(x,y) + iv(x,y)]}{\delta x} \\ &= \lim_{\delta x \to 0} \frac{[u(x+\delta x, y) - u(x,y)] + i[v(x+\delta x, y) - v(x,y)]}{\delta x} \qquad = u_x + iv_x \end{aligned}$$

Similarly, if we assume δz wholly imaginary number, then

$$\begin{aligned} f'(z) &= \lim_{\delta y \to 0} \frac{[u(x, y+\delta y) + iv(x, y+\delta y)] - [u(x,y) + iv(x,y)]}{\delta y} \\ &= \lim_{\delta x \to 0} \frac{[u(x, y+\delta y) - u(x,y)] + i[v(x, y+\delta y) - v(x,y)]}{i\delta y} \qquad = \frac{1}{i} u_y + v_y = v_y - iu_y \end{aligned}$$

Since $f'(z)$ exists, it is unique, therefore

$$u_x + iv_x = v_y - iu_y$$

Equating then the real and imaginary parts, we obtain

$$u_x = v_y \text{ and } u_y = -v_x$$

Thus the necessary conditions for the existence of the derivative of $f(z)$ is that the CR equations should be satisfied.

Sufficient Condition

Suppose $f(z)$ possessing partial derivatives u_x, u_y, v_x, v_y at each point in D and the CR equations are satisfied.

$$\begin{aligned} f(z) &= u(x,y) + iv(x,y) \\ f(z+\delta z) &= u(x+\delta x, y+\delta y) + iv(x+\delta x, y+\delta y) \\ &= [u(x,y) + (u_x\delta x + u_y\delta y) + \ldots] + i[v(x,y) + (v_x\delta x + v_y\delta y) + \ldots] \\ &\qquad \text{(Using Taylor's Theorem for two variables)} \\ &= [u(x,y) + iv(x,y)] + (u_x + iv_x)\delta x + (u_y + iv_y)\delta y + \ldots \\ &= f(z) + (u_x + iv_x)\delta x + (u_y + iv_y)\delta y \\ &= \qquad \text{Leaving the higher order terms} \\ \Rightarrow \quad f(z+\delta z) - f(z) &= (u_x + iv_x)\delta x + (u_y + iv_y)\delta y \end{aligned}$$

On using Cauchy Riemann equations

$$u_x = v_y;\ u_y = -v_x$$

we get,

$$\begin{aligned}
f(z+\delta z) - f(z) &= (u_x + iv_x)\delta x + (-v_x + iu_x)\delta y \\
&= (u_x + iv_x)\delta x + i(iv_x + u_x)\delta y \\
&= (u_x + iv_x)(\delta x + i\delta y) \\
&= (u_x + iv_x)\delta z \\
\Rightarrow \quad \frac{f(z+\delta z) - f(z)}{\delta z} &= u_x + iv_x \\
\Rightarrow \quad \lim_{z\to 0} \frac{f(z+\delta z) - f(z)}{\delta z} &= u_x + iv_x \\
\Rightarrow \quad f'(z) &= u_x + iv_x
\end{aligned}$$

Since u_x, v_x exist and are unique, therefore we conclude that $f'(z)$ exists. Hence $f(z)$ is analytic.

□

Remark : $\dfrac{dw}{dz} = u_x + iv_x = \dfrac{\partial u}{\partial x} + i\dfrac{\partial v}{\partial x} = \dfrac{\partial}{\partial x}(u+iv) = \dfrac{\partial w}{\partial x}$. Also, $\dfrac{dw}{dz} = -i\dfrac{\partial w}{\partial y}$(why?).

■ EXAMPLE 1.5

Discuss the exponential function.

$$w = e^z = \phi(x,y) = e^x(\cos y + i\sin y)$$

Solution: We use the Cauchy-Riemann equations to show that the function is entire. Let

$$f(z) = u + iv = e^z = e^x(\cos y + i\sin y)$$

Hence

$$u = e^x \cos y \text{ and } v = e^x \sin y$$

Then,

$$u_x = e^x \cos y, \quad u_y = -e^x \sin y, \quad v_x = e^x \sin y, \quad v_y = e^x \cos y$$

It follows that Cauchy-Riemann equations are satisfied. Since the function satisfies the Cauchy-Riemann equations and the first partial derivatives are continuous everywhere in the finite complex plane. Hence $f'(z)$ exists and

Now we find the value of the complex derivative.

$$\begin{aligned}
f'(z) &= \frac{dw}{dz} = \frac{\partial w}{\partial x} \\
&= \frac{\partial}{\partial x}[e^x(\cos y + i\sin y)] \\
&= e^x(\cos y + i\sin y) \\
&= e^z
\end{aligned}$$

□

Remark: The differentiability of the exponential function implies the differentiability of the trigonometric functions, as they can be written in terms of the exponential.

■ **EXAMPLE 1.6**

A function $f(z)$ is defined as follows:

$$f(z) = \begin{cases} \frac{x^3-y^3}{x^2+y^2} + i\frac{x^3+y^3}{x^2+y^2}, z \neq 0 \\ 0, \qquad\qquad\qquad\quad z = 0 \end{cases}$$

Show that $f(z)$ is continuous and that Cauchy-Riemann equations are satisfied at the origin. Also show that $f'(0)$ does not exist.

Solution: We have

$$u = \frac{x^3-y^3}{x^2+y^2} \quad \text{and} \quad v = \frac{x^3+y^3}{x^2+y^2}$$

For non zero values of z, $f(z)$ is continuous since u and v are rational functions of x and y with non-zero denominators. To prove its continuity at $z = 0$, we use polar coordinates. Then we have $u = r(\cos^3\theta - \sin^3\theta)$ and $v = r(\cos^3\theta + \sin^3\theta)$. It is seen that u and v tends to zero as $r \to 0$ irrespective of the values of θ. Since $u(0,0) = v(0,0) = 0$ (Given $f(z) = 0$ at $z = 0$), it follows that $f(z)$ is continuous at $(0,0)$. Thus $f(z)$ is continuous for all values of z.

Further,

$$(u_x)_{(0,0)} = \lim_{x\to 0} \frac{u(x,0)-u(0,0)}{x} = \lim_{x\to 0} \frac{x^3/x^2}{x} = 1$$

$$(u_y)_{(0,0)} = \lim_{y\to 0} \frac{u(0,y)-u(0,0)}{y} = \lim_{y\to 0} \frac{-y^3/y^2}{y} = -1$$

$$(v_x)_{(0,0)} = \lim_{x\to 0} \frac{v(x,0)-v(0,0)}{x} = \lim_{x\to 0} \frac{x^3/x^2}{x} = 1$$

$$(v_y)_{(0,0)} = \lim_{y\to 0} \frac{v(0,y)-u(0,0)}{y} = \lim_{y\to 0} \frac{y^3/y^2}{y} = 1$$

which show that the Cauchy-Riemann equations are satisfied at the origin. Finally, Now, let $z \to 0$ along $y = mx$, then

$$\begin{aligned} f'(0) &= \lim_{z\to 0} \frac{f(z)-f(0)}{z} \\ &= \lim_{x,y\to 0} \frac{(x^3-y^3)+i(x^3+y^3)}{(x^2+y^2)(x+iy)} \\ &= \lim_{x\to 0} \frac{x^3(1-m^3)+ix^3(1+m^3)}{x^2(1+m^2)x(1+im)} \\ &= \lim_{x\to 0} \frac{(1-m^3)+i(1+m^3)}{(1+m^2)(1+im)} \end{aligned}$$

Since this limit depends on m therefore $f'(0)$ is not unique, it follows that $f'(z)$ does not exist at $z = 0$. □

1.3 CAUCHY-RIEMANN EQUATION IN POLAR FORM

We know that $x = r\cos\theta, y = r\sin\theta$ and u is a function x and y. Thus, we have

$$z = x + iy = r\cos\theta + ir\sin\theta = re^{i\theta}$$

$$\Rightarrow f(z) = u + iv = f(re^{i\theta}) \tag{1.1}$$

Differentiating Equation (1.1) partially with respect to r, we get

$$\frac{\partial u}{\partial r} + i\frac{\partial v}{\partial v} = f'(re^{i\theta})e^{i\theta} \tag{1.2}$$

Again, differentiating Equation (1.10) partially with respect to θ, we get

$$\frac{\partial u}{\partial \theta} + i\frac{\partial v}{\partial \theta} = if'(re^{i\theta})re^{i\theta} \tag{1.3}$$

Substituting the value of $f'(re^{i\theta})e^{i\theta}$ from Equation (1.2) into Equation (1.3), we get

$$\frac{\partial u}{\partial \theta} + i\frac{\partial v}{\partial \theta} = \left(\frac{\partial u}{\partial r} + i\frac{\partial v}{\partial r}\right) ir$$

or

$$\frac{\partial u}{\partial \theta} + i\frac{\partial v}{\partial \theta} = ir\frac{\partial u}{\partial r} - r\frac{\partial v}{\partial r}$$

Comparing real and imaginary parts of the above equation, we get

$$\frac{\partial v}{\partial r} = -\frac{1}{r}\frac{\partial u}{\partial \theta} \tag{1.4}$$

and

$$\frac{\partial u}{\partial r} = \frac{1}{r}\frac{\partial v}{\partial \theta} \tag{1.5}$$

1.4 DERIVATIVE OF W IN POLAR FORM

We have

$$w = u + iv$$

Therefore

$$\frac{dw}{dz} = \frac{\partial w}{\partial x} = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}$$

But

$$\begin{aligned}\frac{dw}{dz} = \frac{\partial w}{\partial x} &= \frac{\partial w \partial r}{\partial r \partial x} + \frac{\partial w \partial \theta}{\partial \theta \partial x} \\ &= \frac{\partial w}{\partial r}\cos\theta - \left(\frac{\partial u}{\partial \theta} + i\frac{\partial v}{\partial \theta}\right)\frac{\sin\theta}{r} \quad (w = u + iv) \\ &= \frac{\partial u}{\partial r}\cos\theta - \left(-r\frac{\partial v}{\partial r} + ir\frac{\partial u}{\partial r}\right)\frac{\sin\theta}{r}\end{aligned}$$

Since $\frac{\partial u}{\partial \theta} = -r\frac{\partial v}{\partial r}$ and $\frac{\partial v}{\partial \theta} = r\frac{\partial u}{\partial r}$. Therefore

$$\frac{dw}{dz} = \frac{\partial w}{\partial r}\cos\theta - i\left(\frac{\partial u}{\partial r} + i\frac{\partial v}{\partial r}\right)\sin\theta = \frac{\partial w}{\partial r}\cos\theta - i\frac{\partial}{\partial r}(u + iv)\sin\theta$$

$$\therefore \quad \frac{dw}{dz} = (\cos\theta - i\sin\theta)\frac{\partial w}{\partial r} \tag{1.6}$$

Again, we have

$$\frac{dw}{dz} = \frac{\partial w \partial r}{\partial r \partial x} + \frac{\partial w \partial \theta}{\partial \theta \partial x} = \left(\frac{\partial u}{\partial r} + i\frac{\partial v}{\partial r}\right)\cos\theta - \frac{\partial w}{\partial \theta}\frac{\sin\theta}{r}$$

Using the Cauchy-Riemann equations in Polar form, we get

$$= (\frac{1}{r}\frac{\partial v}{\partial \theta} - i\frac{1}{r}\frac{\partial u}{\partial \theta})\cos\theta - \frac{\partial w}{\partial \theta}\frac{\sin\theta}{r} = -\frac{1}{r}(\frac{\partial u}{\partial \theta} + i\frac{\partial v}{\partial \theta})\cos\theta - \frac{\partial w}{\partial \theta}\frac{\sin\theta}{r}$$

$$\therefore \quad \frac{dw}{dz} = -\frac{i}{r}(\cos\theta - i\sin\theta)\frac{\partial w}{\partial \theta} \tag{1.7}$$

1.5 ANALYTIC FUNCTIONS

We will consider a single valued function throughout the section.

1.5.1 Analyticity at point

The function $f(z)$ is said to be analytic at a point $z = z_0$ in the domain D if its derivative $f'(z)$ exists at $z = z_0$ and at every point in some neighborhood of z_0.

1.5.2 Analyticity in domain

A function $f(z)$ is said to be analytic in a domain D, if $f(z)$ is defined and differentiable at all points of the domain.

Note that complex differentiable has a different meaning than analytic. Analyticity refers to the behavior of a function on an open set. A function can be complex differentiable at isolated points, but the function would not be analytic at those points. Analytic functions are also called *regular* or *holomorphic*. If a function is analytic everywhere in the finite complex plane, it is called *entire*.

■ **EXAMPLE 1.7**

Consider z^n, $n \in \mathbb{Z}^+$, Is the function differentiable? Is it analytic? What is the value of the derivative?

Solution: We determine differentiability by trying to differentiate the function. We use the limit definition of differentiation. We will use Newton's binomial formula to expand $(z+\Delta z)^n$.

$$\begin{aligned}
\frac{d}{dz}z^n &= \lim_{\Delta z \to 0} \frac{(z+\Delta z)^n - z^n}{\Delta z} \\
&= \lim_{\Delta z \to 0} \frac{\left(z^n + nz^{n-1}\Delta z + \frac{n(n-1)}{2}z^{n-2}\Delta z^2 + \cdots + \Delta z^n\right) - z^n}{\Delta z} \\
&= \lim_{\Delta z \to 0} \left(nz^{n-1} + \frac{n(n-1)}{2}z^{n-2}\Delta z + \cdots + \Delta z^{n-1}\right) \\
&= nz^{n-1}
\end{aligned}$$

The derivative exists everywhere. The function is analytic in the whole complex plane so it is entire. The value of the derivative is $\frac{d}{dz} = nz^{n-1}$. □

Remark: The definition of the derivative of a function of complex variable is identical in form of that the derivative of the function of real variable. Hence the rule of differentiation for complex functions are the same as those of real functions. Thus if a complex function is once known to be analytic, it can be differentiated just like ordinary way.

■ EXAMPLE 1.8

If $w = \log z$, find $\dfrac{dw}{dz}$ and determine the value of z at which function ceases to be analytic.

Solution: We have

$$w = \log z = \log(x+iy) = \frac{1}{2}\log(x^2+y^2) + i\tan^{-1}\frac{y}{x}$$

i.e.,

$$u = \frac{1}{2}\log(x^2+y^2)$$
$$v = \tan^{-1}\frac{y}{x}$$
$$\therefore \quad u_x = \frac{x}{x^2+y^2}, u_y = \frac{y}{x^2+y^2}$$
$$\text{and} \quad v_x = \frac{-y}{x^2+y^2}, v_y = \frac{x}{x^2+y^2}$$

Since, the CR equations are satisfied and the partial derivatives are continuous except at (0,0). Hence w is analytic everywhere except at $z = 0$.

$$\therefore \quad \frac{dw}{dz} = u_x + iv_x = \frac{x}{x^2+y^2} + i\frac{-y}{x^2+y^2} = \frac{x-iy}{x^2+y^2} = \frac{1}{x+iy} = \frac{1}{z}$$

where $z \neq 0$. (Note direct differentiations of $\log z$ also gives $1/z$). □

■ EXAMPLE 1.9

Show that for the analytic function $f(z) = u + iv$, the two families of curves $u(x,y) = c_1$ and $v(x,y) = c_2$ are orthogonal[2].

Solution: Families of curves

$$u(x,y) = c_1 \tag{1.8}$$
$$v(x,y) = c_2 \tag{1.9}$$

On differentiating equation 1.8,

$$u_x dx + u_y dy = 0 \qquad \Rightarrow \qquad m_1 = \frac{dy}{dx} = -\frac{u_x}{u_y} \tag{1.10}$$

On differentiating equation 1.8,

$$v_x dx + v_y dy = 0 \qquad \Rightarrow \qquad m_2 = \frac{dy}{dx} = -\frac{v_x}{v_y} \tag{1.11}$$

[2]Two curves are said to be orthogonal if they intersect at right angle at each point of intersection. Mathematically, if the curves have slopes m_1 and m_2, then the curves are orthogonal if $m_1 m_2 = -1$.

The product of two slopes

$$m_1 m_2 = \left(-\frac{u_x}{u_y}\right)\left(-\frac{v_x}{v_y}\right) \tag{1.12}$$

Since, $u+iv$ is analytic, Hence. Cauchy Riemann equations are

$$u_x = v_y \quad \textit{and} \quad u_y = -v_x$$

Hence equation 1.12 reduces to

$$m_1 m_2 = \left(-\frac{u_x}{u_y}\right)\left(\frac{u_y}{u_x}\right) = -1$$

Hence the two families of curves $u(x,y) = c_1$ and $v(x,y) = c_2$ are orthogonal. □

1.5.3 Analytic Functions can be Written in Terms of z.

Consider an analytic function expressed in terms of x and y, $\phi(x,y)$. We can write ϕ as a function of $z = x+iy$ and $\bar{z} = x-iy$.

$$f(z,\bar{z}) = \phi\left(\frac{z+\bar{z}}{2}, \frac{z-\bar{z}}{2i}\right)$$

We treat z and $\bar{z}$ as independent variables. We find the partial derivatives with respect to these variables.

$$\frac{\partial}{\partial z} = \frac{\partial x}{\partial z}\frac{\partial}{\partial x} + \frac{\partial y}{\partial z}\frac{\partial}{\partial y} = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right)$$
$$\frac{\partial}{\partial \bar{z}} = \frac{\partial x}{\partial \bar{z}}\frac{\partial}{\partial x} + \frac{\partial y}{\partial \bar{z}}\frac{\partial}{\partial y} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)$$

Since ϕ is analytic, the complex derivatives in the x and y directions are equal.

$$\frac{\partial \phi}{\partial x} = -i\frac{\partial \phi}{\partial y}$$

The partial derivative of $f(z,\bar{z})$ with respect to $\bar{z}$ is zero.

$$\frac{\partial f}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial \phi}{\partial x} + i\frac{\partial \phi}{\partial y}\right) = 0$$

Thus $f(z,\bar{z})$ has no functional dependence on $\bar{z}$, it can be written as a function of z alone.

If we were considering an analytic function expressed in polar coordinates $\phi(r,\theta)$, then we could write it in Cartesian coordinates with the substitutions:

$$r = \sqrt{x^2+y^2}, \quad \theta = \arctan(x,y).$$

Thus we could write $\phi(r,\theta)$ as a function of z alone.

■ **EXAMPLE 1.10**

If n is real, show that $f(re^{i\theta}) = r^n(\cos n\theta + i\sin n\theta)$ is analytic expect possibly when $r = 0$ and that its derivative is $nr^{n-1}[\cos(n-1)\theta + i\sin(n-1)\theta]$.

Solution: Let $w = f(z) = u+iv = r^n(\cos n\theta + i\sin n\theta)$. Therefore $u = r^n\cos n\theta, v = r^n\sin n\theta$

$$\frac{\partial u}{\partial r} = nr^{n-1}\cos n\theta, \frac{\partial u}{\partial \theta} = -nr^n \sin n\theta$$

$$\frac{\partial v}{\partial r} = nr^{n-1}\sin n\theta, \frac{\partial v}{\partial \theta} = nr^n \cos n\theta$$

Thus, we have

$$\frac{\partial u}{\partial r} = \frac{1}{r}\frac{\partial v}{\partial \theta} = nr^{n-1}\cos n\theta$$

and

$$\frac{1}{r}\frac{\partial u}{\partial \theta} = -\frac{\partial v}{\partial r} = -nr^{n-1}\sin n\theta$$

Hence, the Cauchy-Riemann equations are satisfied. Thus, the function $w = r^n(\cos n\theta + i\sin n\theta)$ is analytic for all finite values of z, if $\frac{dw}{dz}$ exits . we have

$$\frac{dw}{dz} = (\cos\theta - i\sin\theta)\frac{\partial w}{\partial r} = (\cos\theta - i\sin\theta)\frac{\partial w}{\partial r} = (\cos\theta - i\sin\theta)nr^{n-1}(\cos n\theta + i\sin n\theta)$$

$$= nr^{n-1}[\cos(n-1)\theta + i\sin(n-1)\theta]$$

Thus, $\dfrac{dw}{dz}$ exits for all values of r, including zero, expect when $r = 0$ and $n \leq 1$. □

■ EXAMPLE 1.11

Show that the function $f(z) = e^{-z^{-4}}$ $(z \neq 0)$ and $f(0) = 0$ is not analytic at $z = 0$. Although Cauchy-Riemann equations are satisfied at the point. How would you explain this?

Solution: Here

$$\begin{aligned} f(z) &= e^{-z^{-4}} \\ &= e^{-\frac{1}{(x+iy)^4}} = e^{-\frac{(x-iy)^4}{(x^2+y^2)^4}} = e^{-\frac{(x^4+y^4-6x^2y^2)-i4xy(x^2-y^2)}{(x^2+y^2)^4}} \\ \Rightarrow u+iv &= e^{-\frac{x^4+y^4-6x^2y^2}{(x^2+y^2)^4}} e^{-i\frac{4xy(x^2-y^2)}{(x^2+y^2)^4}} \end{aligned}$$

This gives,

$$u = e^{-\frac{x^4+y^4-6x^2y^2}{(x^2+y^2)^4}} \cos\left(\frac{4xy(x^2-y^2)}{(x^2+y^2)^4}\right) \text{ and } v = e^{-\frac{x^4+y^4-6x^2y^2}{(x^2+y^2)^4}} \sin\left(\frac{4xy(x^2-y^2)}{(x^2+y^2)^4}\right)$$

At $z = 0$

$$\begin{aligned} \frac{\partial u}{\partial x} &= \lim_{h\to 0}\frac{u(0+h,0)-u(0,0)}{h} = \lim_{h\to 0}\frac{e^{-h^{-4}}}{h} \\ &= \lim_{h\to 0}\frac{1}{he^{\frac{1}{h^4}}} = \lim_{h\to 0}\left[\frac{1}{h\left(1+\frac{1}{h^4}+\frac{1}{2!h^8}+\frac{1}{3!h^{12}}+\ldots\right)}\right] = 0 \\ &\left(\because e^x = 1+x+\frac{x^2}{2!}+\ldots\right) \end{aligned}$$

$$\frac{\partial u}{\partial y} = \lim_{k\to 0} \frac{u(0,0+k)-u(0,0)}{k} = \lim_{k\to 0} \frac{e^{-k^{-4}}}{k}$$
$$= \lim_{k\to 0} \frac{1}{he^{\frac{1}{h^4}}} = 0$$

$$\frac{\partial v}{\partial x} = \lim_{h\to 0} \frac{v(0+h,0)-v(0,0)}{h} = \lim_{h\to 0} \frac{e^{-h^{-4}}}{h}$$
$$= \lim_{h\to 0} \frac{1}{h.e^{\frac{1}{h^4}}} = 0$$

$$\frac{\partial v}{\partial y} = \lim_{k\to 0} \frac{v(0,0+k)-v(0,0)}{k} = \lim_{k\to 0} \frac{e^{-k^{-4}}}{k}$$
$$= \lim_{k\to 0} \frac{1}{k.e^{\frac{1}{k^4}}} = \lim_{k\to 0} \frac{1}{k.e^{\frac{1}{k^4}}} = 0$$

Hence $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$ and $\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$ (C-R equations are satisfied at $z=0$)
But

$$f'(0) = \lim_{z\to 0} \frac{f(z)-f(0)}{z} = \lim_{z\to 0} \frac{e^{-z^{-4}}-0}{z}$$

Along $z = re^{i\frac{\pi}{4}}$

$$f'(0) = \lim_{r\to 0} \frac{e^{-r^{-4}} e^{-\left(e^{i\frac{\pi}{4}}\right)^4}}{re^{i\frac{\pi}{4}}} = \lim_{r\to 0} \frac{e^{-r^{-4}} e}{re^{i\frac{\pi}{4}}}$$
$$= \frac{e}{e^{i\frac{\pi}{4}}} \lim_{r\to 0} \frac{1}{re^{-r^{-4}}} = 0$$

Showing that $f'(z)$ does not exist at $z=0$. Hence $f(z)$ is not analytic at $z=0$. □

Problems

1.1 Determine which of the following functions are analytic:

a) $x^2 + iy^2$

b) $2xy + i(x^2 - y^2)$

c) $\sin x \cosh y + i \cos x \sinh y$

d) $\frac{1}{(z-1)(z+1)}$

e) $\frac{x-iy}{x-iy+a}$

f) $\frac{x-iy}{x^2+y^2}$

1.2 Consider the function $f(z) = (4x+y) + i(-x+4y)$ and discuss $\frac{df}{dz}$

1.3 For what values of z, the function w defined as

$$w = \rho(\cos\phi + i\sin\phi); \quad \text{where } z = \ln\rho + i\phi$$

cases to be analytic.

1.4 For what values of z the function $z = \sinh u \cos v + i \cosh u \sin v$, where $w = u + iv$ ceases to be analytic.

1.5 For what values of z the function $z = e^{-v}(\cos u + i \sin u)$, where $w = u + iv$ ceases to be analytic.

1.6 If

$$f(z) = \begin{cases} \dfrac{x^3y(y-ix)}{x^6+y^2}, & z \neq 0 \\ = 0, z = 0, \end{cases}$$

then discuss $\dfrac{df}{dz}$ at $z = 0$.

1.7 Show that the complex variable function $f(z) = |z|^2$ is differentiable only at the origin.

1.8 Using the Cauchy-Riemann equations, show that $f(z) = z^3$ is analytic in the entire z-plane.

1.9 Test the analyticity of the function $w = \sin z$ and hence derive that:

$$\frac{d}{dz}(\sin z) = \cos z$$

1.10 Find the point where the Cauchy-Riemann equations are satisfied for the function:

$$f(z) = xy^2 + ix^2y$$

where does $f'(z)$ = exist? Where is $f(z)$ analytic?

1.11 Find the values of a and b such that the function

$$f(z) = x^2 + ay^2 - 2xy + i(bx^2 - y^2 + 2xy)$$

is analytic. Also find $f'(z)$.

1.12 Show that the function $z|z|$ is not analytic anywhere.

1.13 Discuss the analyticity of the function $f(z) = z\bar{(z)}$.

1.14 Show that the function $f(z) = u + iv$, where

$$f(z) = \begin{cases} \dfrac{x^3(1+i) - y^3(1-i)}{x^2+y^2}, & z \neq 0 \\ \text{=0, z=0} \end{cases}$$

satisfy the Cauchy-Riemann conditions at $z = 0$. Is the function analytic at $z = 0$? justify your answer.

1.15 Show that the function defined by $f(z) = \sqrt{|xy|}$ satisfy Cauchy Riemann equations at the origin but is not analytic at the point.

1.6 HARMONIC FUNCTIONS

Any real valued function of x and y satisfying Laplace equation[3] is called Harmonic function.
If $f(z) = u + iv$ is analytic function then u and v are harmonic functions.
If $f(z)$ is analytic, we have CR Equations,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

On differentiating first equation partially with respect to x and second equation partially with respect to y, we get

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial y \partial x} \qquad \frac{\partial^2 u}{\partial y^2} = -\frac{\partial^2 v}{\partial y \partial x}$$

On adding both equations, we get

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

which shows that u is harmonic. Similarly, on differentiating first equation partially with respect to y and second equation partially with respect to x, we get

$$\frac{\partial^2 u}{\partial y \partial x} = \frac{\partial^2 v}{\partial x^2} \qquad \frac{\partial^2 u}{\partial y \partial x} = -\frac{\partial^2 v}{\partial x^2}$$

On subtracting both equations, we get

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0$$

which shows that v is harmonic. Hence, if $f(z) = u + iv$ is some analytic function then u and v are harmonic functions.

1.7 DETERMINATION OF CONJUGATE FUNCTIONS

If $f(z) = u + iv$ is an analytic function, $v(x,y)$ is called conjugate function of $u(x,y)$. In this section we have to device a method to compute $v(x,y)$ provided $u(x,y)$ is given From partial differentiation, we have

$$dv = \frac{\partial v}{\partial x} dx + \frac{\partial v}{\partial y} dy \tag{1.13}$$

But, $f(z)$ is analytic, which implies

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \tag{1.14}$$

[3]The following equation is known as Laplace equation

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

Using equations 1.14, Equation 1.13 reduces to

$$dv = -\frac{\partial u}{\partial y}dx + \frac{\partial u}{\partial x}dy \tag{1.15}$$

Here $M = -\dfrac{\partial u}{\partial y}$ and $N = \dfrac{\partial u}{\partial x}$, which gives

$$\frac{\partial M}{\partial y} = -\frac{\partial^2 u}{\partial y^2} \qquad \frac{\partial N}{\partial x} = \frac{\partial^2 u}{\partial x^2}$$

$$\frac{\partial M}{\partial y} - \frac{\partial N}{\partial y} = -\left(\frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial x^2}\right)$$

Since $f(z)$ is analytic, u is harmonic

$$\frac{\partial M}{\partial y} - \frac{\partial N}{\partial y} = 0$$

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial y}$$

which shows that Eq 1.15 is an exact differential equation. It can be integrated to obtain v.

■ EXAMPLE 1.12

Show that $u = x^2 - y^2$ is harmonic and find the its conjugate

Solution: Here u is given, we may compute following

$$u_x = 2x \qquad u_y = -2y$$

$$u_{xx} = 2 \qquad u_{yy} = -2$$

This implies $u_{xx} + u_y y = 0$, i.e., Laplace equation holds. Therefore, the given function is harmonic. From partial differentiation and CR Equations, we have

$$dv = -\frac{\partial u}{\partial y}dx + \frac{\partial u}{\partial x}dy$$

which gives

$$dv = 2ydx + 2xdy$$

As this differential equation is exact, we may use method of solving an exact differential equation[4], which is as

$$v = \int_{\text{y as constant}} (2y)dx + c$$

$$v = 2xy + c$$

which is required harmonic conjugate of u. □

[4]See Appendix

■ **EXAMPLE 1.13**

If ϕ and ψ are function of x and y satisfying Laplace's equation, show that $s+it$ is analytic, where
$s = \frac{\partial\phi}{\partial y} - \frac{\partial\psi}{\partial x}$ and $t = \frac{\partial\phi}{\partial x} + \frac{\partial\psi}{\partial y}$.

Solution:
Since ϕ and ψ are function of x and y satisfying Laplace's equations.

$$\therefore \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} = 0 \tag{1.16}$$

and

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial\psi}{\partial y^2} = 0. \tag{1.17}$$

For the function $s+it$ to be analytic,

$$\frac{\partial s}{\partial x} = \frac{\partial t}{\partial y} \tag{1.18}$$

$$\frac{\partial s}{\partial y} = -\frac{\partial t}{\partial x} \tag{1.19}$$

must satisfy.
Now,

$$\frac{\partial s}{\partial x} = \frac{\partial}{\partial x}\left(\frac{\partial\phi}{\partial y} - \frac{\partial\psi}{\partial x}\right) = \frac{\partial^2\phi}{\partial x\partial y} - \frac{\partial^2\psi}{\partial x^2} \tag{1.20}$$

$$\frac{\partial t}{\partial y} = \frac{\partial}{\partial y}\left(\frac{\partial\phi}{\partial x} + \frac{\partial\psi}{\partial y}\right) = \frac{\partial^2\phi}{\partial y\partial x} + \frac{\partial^2\psi}{\partial y^2} \tag{1.21}$$

$$\frac{\partial s}{\partial y} = \frac{\partial}{\partial y}\left(\frac{\partial\phi}{\partial y} - \frac{\partial\psi}{\partial x}\right) = \frac{\partial^2\phi}{\partial y^2} - \frac{\partial^2\psi}{\partial y\partial x} \tag{1.22}$$

and

$$\frac{\partial t}{\partial x} = \frac{\partial}{\partial x}\left(\frac{\partial\phi}{\partial x} + \frac{\partial\psi}{\partial y}\right) = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\psi}{\partial x\partial y} \tag{1.23}$$

From (1.18), (1.20) and (1.21), we have

$$\frac{\partial^2\phi}{\partial x\partial y} - \frac{\partial^2\psi}{\partial x^2} = \frac{\partial^2\phi}{\partial y\partial x} + \frac{\partial^2\psi}{\partial y^2}$$

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} = 0$$

Which is true by (1.17).
Again from (1.19), (1.22) and (1.23), we have

$$\frac{\partial^2\phi}{\partial y^2} - \frac{\partial^2\psi}{\partial y\partial x} = -\frac{\partial^2\phi}{\partial x^2} - \frac{\partial^2\psi}{\partial x\partial y}$$

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} = 0$$

which is also true by (1.16).
Hence the function $s+it$ is analytic.

□

■ **EXAMPLE 1.14**

Show that an analytic function with constant modulus is constant.

Solution:

Solution : Let $f(z) = u + iv$ be an analytic function with constant modulus. Then,

$$|f(z)| = |u + iv| = \text{Constant}$$

$$\sqrt{u^2 + v^2} = \text{Constant} = c \text{ (say)}$$

Squaring both sides, we get

$$u^2 + v^2 = c^2 \tag{1.24}$$

Differentiating equation (1.24) partially w.r.t. x, we get

$$2u\frac{\partial u}{\partial x} + 2v\frac{\partial v}{\partial x} = 0$$

$$u\frac{\partial u}{\partial x} + v\frac{\partial v}{\partial x} = 0 \tag{1.25}$$

Again, differentiating equation (1.24) partially w.r.t. y, we get

$$2u\frac{\partial u}{\partial y} + 2v\frac{\partial v}{\partial y} = 0$$

$$u\frac{\partial u}{\partial y} + v\frac{\partial v}{\partial y} = 0$$

$$-u\frac{\partial v}{\partial x} + v\frac{\partial u}{\partial x} = 0 \tag{1.26}$$

$\because \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$ and $\frac{\partial v}{\partial y} = \frac{\partial u}{\partial x}$. Squaring and adding equations (1.25) and (1.26), we get

$$(u^2 + v^2)\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2\right] = 0$$

$$\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2 = 0 \quad \because u^2 + v^2 = c^2 \neq 0$$

$$|f'(z)|^2 = 0 \quad [\because f'(z) = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}]$$

$$|f'(z)| = 0$$

Hence $f(z)$ is constant.

□

1.8 MILNE THOMSON METHOD

Consider the problem to determine the function $f(z)$ of which u is given. One procedure may be as previous section, compute its harmonic conjugate v, and finally combine them to compute $f(z) = u + iv$. To overcome the length of the mechanism Milne's introduced another method which is a direct way to compute $f(z)$ for a given u. We have $z = x + iy$

which implies

$$x = \frac{z+\bar{z}}{2} \qquad y = \frac{z-\bar{z}}{2i}$$

$$w = f(z) = u + iv = u(x,y) + iv(x,y)$$

$$= u\left(\frac{z+\bar{z}}{2}, \frac{z-\bar{z}}{2i}\right) + iv\left(\frac{z+\bar{z}}{2}, \frac{z-\bar{z}}{2i}\right)$$

On putting $z = \bar{z}$, we get

$$f(z) = u(z,0) + iv(z,0)$$

We have (CR Equation are used.)

$$f'(z) = \frac{dw}{dz} = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = \frac{\partial u}{\partial x} - i\frac{\partial u}{\partial y}$$

Here, $\dfrac{\partial u}{\partial x} = \phi_1(x,y)$ and $\dfrac{\partial u}{\partial y} = \phi_2(x,y)$ Thus

$$f'(z) = \phi_1(x,y) - i\phi_2(x,y)$$

or

$$f'(z) = \phi_1(z,0) - i\phi_2(z,0)$$

On integrating, we obtain the required function

$$f(z) = \int [\phi_1(z,0) - i\phi_2(z,0)]dz + K$$

where K is complex constant.

Remark: In case, v is given, $\phi_1(x,y) = u_x = v_y$ and $\phi_2(x,y) = u_y = -v_x$

■ EXAMPLE 1.15

Find the regular[5] function. for the given $u = x^2 - y^2$.

Solution: Here u is given, we may compute following

$$\phi_1(x,y) = \frac{\partial u}{\partial x} = 2x \qquad \phi_2(x,y) = \frac{\partial u}{\partial y} = -2y$$

$$\begin{aligned} f'(z) &= \phi_1(z,0) - i\phi_2(z,0) \\ &= 2z - i(0) \\ &= 2z \end{aligned}$$

On integrating,

$$f(z) = 2\int z dz = z^2 + K$$

which is required function. □

[5] Analytic function is also called as Regular function.

EXAMPLE 1.16

If $u-v=(x-y)(x^2+4xy+y^2)$ and $f(z)=u+iv$ is an analytic function of $z=x+iy$, find the $f(z)$ in terms of z.

Solution: We have,

$$u+iv=f(z)$$

This gives

$$iu-v=if(z)$$

On adding these both

$$(u-v)+i(u+v)=(1+i)f(z)$$

Let $U=u-v$ and $V=u+v$ then

$$U+iV=(1+i)f(z)=F(z)\ \ (say)$$

Here $U=(u-v)$ gives,

$$U=(x-y)(x^2+4xy+y^2)$$
$$U=x^3+3x^2y-3xy^2-y^3$$

Now, we can use Milne's Method to find $F(z)$.

$$\phi_1(x,y)=\frac{\partial U}{\partial x}=3x^2+6xy-3y^2 \qquad \phi_2(x,y)=\frac{\partial U}{\partial y}=3x^2-6xy-3y^2$$

$$F'(z)=\phi_1(z,0)-i\phi_2(z,0)$$
$$=3z^2-i3z^2=(1-i)3z^2$$

Hence

$$F(z)=(1-i)z^3+K$$

where K is complex constant.
Since we have $F(z)=(1+i)f(z)$,

$$f(z)=\frac{F(z)}{(1+i)}=\frac{(1-i)z^3+K}{1+i}$$

$$f(z)=-iz^3+K_1$$

where K_1 is complex constant. □

Problems

1.16 Show that the following functions are harmonic and determine the conjugate functions.

- **a)** $u=2x(1-y)$
- **b)** $u=2x-x^3+3xy$
- **c)** $u=\frac{1}{2}log(x^2+y^2)$
- **d)** $x^3-3xy^2+3x^2-3y^2$

1.17 Determine the analytic function, whose imaginary part is

- **a)** $x^2-y^2+5x+y-\frac{y}{x^2+y^2}$

b) $\cos x \cosh y$
c) $3x^2y + 2x^2 - y^3 - 2y^2$
d) $e^{-x}(x\sin y - y\cos y)$
e) $e^{2x}(x\cos 2y - y\sin 2y)$
f) $v = \log(x^2 + y^2) + x - 2y$
g) $v = \sinh x cosy$
h) $v = \frac{x-y}{x^2+y^2}$
i) $v = \left(r - \frac{1}{r}\right) sin\theta$

1.18 If $f(z) = u + iv$ is an analytic function of $z = x + iy$ and $u - v = \dfrac{e^y - \cos x + \sin x}{\cosh y - \cos x}$, find $f(z)$ subject to the condition that $f\left(\frac{\pi}{2}\right) = \frac{(3-i)}{2}$

1.19 Find an analytic function $f(z) = u(r,\theta) + iv(r,\theta)$ such that $v(r,\theta) = r^2\cos 2\theta - r\cos\theta + 2$

1.20 Show that the function $u = x^2 - y^2 - 2xy - 2x - y - 1$ is harmonic. Find the conjugate harmonic function v and express $u + iv$ as a function of z where $z = x + iy$

1.21 Construct an analytic function of the form $f(z) = u + iv$, where v is $tan^{-1}(\frac{y}{x})$, $x \neq 0, y \neq 0$

1.22 If $f(z)$ is a regular function of z, prove that $\left(\dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}\right)|f(z)|^2 = 4|f'(z)|^2$

CHAPTER 2

COMPLEX INTEGRATION

2.1 INTEGRATION IN COMPLEX PLAIN

In case of real variable, the path of the integration of $\int_a^b f(x)dx$ is always along the x-axis from $x=a$ to $x=b$. But in case of a complex function $f(z)$ the path of a complex function $f(z)$ the path of the definite integral $\int_\alpha^\beta f(z)dz$ can be along any curve from $z=\alpha$ to $z=\beta$.

■ EXAMPLE 2.1

Evaluate $\int_0^{2+i} \bar{z}^2 dz$ along the real axis from $z=0$ to $z=2$ and then along parallel to y-axis from $z=2$ to $z=2+i$.

Solution:

$$\begin{aligned}\int_0^{2+i} \bar{z}^2 dz &= \int_0^{2+i} (x-iy)^2(dx+idy)\\ &= \int_0^{2+i} (x^2-y^2-2ixy)(dx+idy)\end{aligned}$$

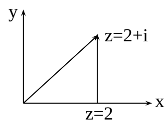


Figure 2.1.

Along real axis from $z=0$ to $z=2$ (y=0). :

$$y = 0 \Rightarrow dy = 0, dz = d(x+iy) = dx$$

$$z = 0, y = 0 \Rightarrow x = 0$$

and

$$z = 2, y = 0 \Rightarrow x = 2$$

$$\int_0^{2+i} \bar{z}^2 dz = \int_0^2 (x^2)(dx)$$
$$= \left[\frac{x^3}{3}\right]_0^2 = \frac{8}{3}$$

Along parallel to y-axis from $z=2$ to $z=2+i$ (x=2).

$$x = 2 \Rightarrow dx = 0, dz = d(x+iy) = idy$$

$$z = 2, x = 2 \Rightarrow y = 0 \quad and \quad z = 2+i, x = 2 \Rightarrow y = 1$$

$$\int_0^{2+i} \bar{z}^2 dz = \int_0^1 (4 - y^2 - 4iy)(i.dy)$$
$$= i\left[4y - \frac{y^3}{3} - 4i\frac{y^2}{2}\right]_0^1 = \left[\frac{11}{3}i + 2\right]$$

$\int_0^{2+i} \bar{z}^2 dz$ along the real axis from $z=0$ to $z=2$ then along parallel to y-axis from $z=2$ to $z=2+i$

$$= \frac{8}{3} + \frac{11}{3}i + 2 = \frac{1}{3}(14 + 11i)$$

□

Problems

2.1 Find the value of the integral

$$\int_0^{1+i} (x - y + ix^2) dz$$

a) Along the straight line from $z=0$ to $z=1+i$.

b) along the real axis from $z=0$ to $z=1$ and then along parallel to y-axis from $z=1$ to $z=1+i$.

2.2 Integrate $f(z)=x^2+ixy$ from $A(1,1)$ to $B(2,8)$ along

a) the straight line AB

b) the curve C, $x=t,\ y=t^3$.

2.3 Evaluate the integral $\int_c (3y^2dx+2ydy)$, where c is the circle $x^2+y^2=1$, counter-clockwise from $(1,0)$ to $(0,1)$.

2.2 CAUCHY'S INTEGRAL THEOREM

Theorem 2. *If a function $f(z)$ is analytic and its derivative $f'(z)$ continuous at all points within and on a simple closed curve c, then* $\int_c f(z)dz=0$.

Proof. Let $f(z)=u+iv$ and $z=x+iy$ and region enclosed by the curve c be R, then

$$\int_c f(z)dz=\int_c (u+iv)(dx+idy)=\int_c (udx-vdy)+i(vdx+udy)$$

$$=\int\int_R (-v_x-u_y)dxdy+i\int\int_R (u_x-v_y)dxdy$$

By Cauchy-Riemann equations,

$$=\int\int_R (u_y-u_y)dxdy+i\int\int_R (u_x-u_x)dxdy=0$$

□

■ **EXAMPLE 2.2**

Find the integral $\int_c \frac{3z^2+7z+1}{z+1}dz$, where c is the circle $|z|=\frac{1}{2}$.

Solution: Poles of integrand are given by

$$z+1=0$$

That is, $z=-1$. Since given circle $|z|=\frac{1}{2}$, with centre $z=0$ and radius $1/2$ does not enclose $z=-1$. Thus it is obvious that the integrand is analytic everywhere. Hence, by Cauchy's Theorem,

$$\int_c \frac{3z^2+7z+1}{z+1}dz=0$$

□

Theorem 3 (Cauchy's integral theorem for multi-connected region)**.** *If a function $f(z)$ is analytic in region R between two simple closed curves c_1 and c_2, then*

$$\int_{c_1} f(z)dz=\int_{c_2} f(z)dz$$

Proof. Since $f(z)$ is analytic in region R, By Cauchy's Theorem

$$\int f(z)dz=0$$

where path of integration is along AB, and curves C_2 in clockwise direction and along BA and along C_1 in anticlockwise direction.

We may write,

$$\int_{AB} f(z)dz - \int_{c_2} f(z)dz + \int_{BA} f(z)dz + \int_{c_1} f(z)dz = 0$$

or

$$-\int_{c_2} f(z)dz + \int_{c_1} f(z)dz = 0$$

$$\int_{c_1} f(z)dz = \int_{c_2} f(z)dz$$

□

2.3 CAUCHY INTEGRAL FORMULA

Theorem 4. *If a function $f(z)$ is analytic within and on a closed curve c, and if a is any point within c , then*

$$f(a) = \frac{1}{2\pi i}\int_{z} \frac{f(z)}{(z-a)}dz$$

Proof. Let $z = a$ be a point within a closed curve c. Describe a circle γ such that $|z-a| = \rho$ and it lies entirely within c. Now consider the function

$$\phi(z) = \frac{f(z)}{(z-a)}$$

Obviously, this function is analytic in region between γ and c. Hence by Cauchy's integral theorem for multiconnected region, we have

$$\int_{c} \phi(z)dz = \int_{\gamma} \phi(z)dz$$

or

$$\begin{aligned}
\int_{c} \frac{f(z)}{(z-a)}dz &= \int_{\gamma} \frac{f(z)}{(z-a)}dz \\
&= \int_{\gamma} \frac{f(z)-f(a)+f(a)}{(z-a)}dz \\
&= \int_{\gamma} \frac{f(z)-f(a)}{(z-a)}dz + \int_{\gamma} \frac{f(a)}{(z-a)}dz \\
&= I_1 + I_2
\end{aligned}$$

Now, since $|z-a| = \rho$, we have $z = a+\rho e^{i\theta}$ and $dz = i\rho e^{i\theta}d\theta$. Hence

$$\begin{aligned}
I_1 &= \int_{\gamma} \frac{f(z)-f(a)}{(z-a)}dz \\
&= \int_{0}^{2\pi} \frac{f(a+\rho e^{i\theta})-f(a)}{[(a+\rho e^{i\theta})-a]} i\rho e^{i\theta}d\theta \\
&= \int_{0}^{2\pi} [f(a+\rho e^{i\theta})-f(a)]id\theta \\
&= 0 \qquad \text{as } \rho \text{ tends to } 0
\end{aligned}$$

and

$$\begin{aligned} I_2 &= \int_\gamma \frac{f(a)}{(z-a)} dz \\ &= \int_0^{2\pi} \frac{f(a)}{[(a+\rho e^{i\theta})-a]} i\rho e^{i\theta} d\theta \\ &= f(a) \int_0^{2\pi} i d\theta \\ &= 2\pi i f(a) \end{aligned}$$

Hence,

$$\int_c \frac{f(z)}{(z-a)} dz = I_1 + I_2$$

That is

$$\int_c \frac{f(z)}{(z-a)} dz = 0 + 2\pi i f(a)$$

or

$$f(a) = \frac{1}{2\pi i} \int_c \frac{f(z)}{(z-a)} dz$$

□

EXAMPLE 2.3

Evaluate (i) $\int_c \frac{e^z}{z+2} dz$ and (ii) $\int_c \frac{e^z}{z} dz$, where c is circle $|z| = 1$.

Solution: (i) The function $\frac{e^z}{z+2}$ is analytic everywhere except at $z = -2$. This point lies outside the circle $|z| = 1$. Thus function is analytic within and on c, by Cauchy's Theorem, we have

$$\int_{|z|=1} \frac{e^z}{z+2} dz = 0$$

(ii) The function $\frac{e^z}{z}$ is analytic everywhere except at $z = 0$. The point $z = 0$ strictly inside $|z| = 1$. Hence by Cauchy's Integral formula, we have

$$\int_{|z|=1} \frac{e^z}{z} dz = 2\pi i (e^z)_{z=0} = 2\pi i$$

□

2.4 CAUCHY INTEGRAL FORMULA FOR THE DERIVATIVES OF AN ANALYTIC FUNCTION

Theorem 5. *If a function $f(z)$ is analytic within and on a closed curve c, and if a is any point within c , then its derivative is also analytic within and on closed curve c, and is given as*

$$f'(a) = \frac{1}{2\pi i} \int_c \frac{f(z)}{(z-a)^2} dz$$

Proof. We know Cauchy's Integral formula

$$f(a)=\frac{1}{2\pi}\int_c \frac{f(z)}{(z-a)}dz$$

Differentiating, wrt a , we get

$$\begin{aligned}f'(a)&=\frac{1}{2\pi}\frac{d}{da}\left[\int_c \frac{f(z)}{(z-a)}dz\right]\\&=\frac{1}{2\pi}\int_c f(z)\frac{\partial}{\partial a}\left[\frac{1}{(z-a)}\right]dz\\&=\frac{1}{2\pi}\int_c \frac{f(z)}{(z-a)^2}dz\end{aligned}$$

We may generalize it,

$$f^n(a)=\frac{n!}{2\pi i}\int_c \frac{f(z)}{(z-a)^{n+1}}dz$$

□

■ EXAMPLE 2.4

Evaluate the following integral $\int_c \frac{1}{z}\cos z dz$, where c is the ellipse $9x^2+4y^2=1$.

Solution: Here function $\frac{1}{z}\cos z$ has a simple pole at $z=0$. The given ellipse $9x^2+4y^2=1$. encloses pole $z=0$.
By Cauchy Integral formula

$$\int_c \frac{cosz}{z}dz=2\pi i(cosz)_{z=0}=2\pi i$$

□

■ EXAMPLE 2.5

Evaluate the complex integral $\int_c \tan z dz$, where c is $|z|=2$.

Solution: We have

$$\int_c tanz.dz=\int_c \frac{\sin z}{\cos z}dz$$

$|z|=2$, is a circle with centre at origin and radius = 2. Poles are given by putting the denominator equal to zero. i.e.,

$$\cos z=0 \quad\Rightarrow\quad z=-\frac{\pi}{2},\frac{\pi}{2},\frac{3\pi}{2},\ldots$$

The integrand has two poles at $z=\frac{\pi}{2}$ and $z=-\frac{\pi}{2}$ inside the given circle $|z|=2$.
On applying Cauchy integral formula

$$\begin{aligned}\int_c \frac{\sin z}{\cos z}dz&=\int_{c_1}\frac{\sin z}{\cos z}dz+\int_{c_2}\frac{\sin z}{\cos z}dz\\&=2\pi i[\sin z]_{z=\frac{\pi}{2}}+2\pi i[\sin z]_{z=-\frac{\pi}{2}}\\&=2\pi i(1)+2\pi i(-1)=0\end{aligned}$$

□

EXAMPLE 2.6

Evaluate $\int_c \frac{e^z}{z^2+1} dz$ over the circular path $|z| = 2$.

Solution: Here,

$$z^2+1=0,\ z^2=-1,\ z=\pm i$$

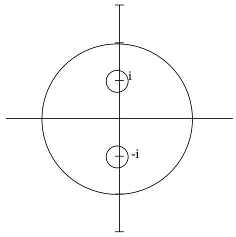


Figure 2.2.

Both points are inside the given circle with at origin and radius 2.

$$\begin{aligned}\int_c \frac{1}{2i}\{\frac{e^z}{z-i} - \frac{e^z}{z+i}\}dz &= \int_c \frac{1}{2i}\frac{e^z}{z-i}dz - \frac{1}{2i}\int_c \frac{e}{z+i}dz \\ &= \frac{1}{2i}[2\pi i(e^z)_{z=i} - 2\pi i(e^z)_{z=-i}] \\ &= \frac{2\pi i}{2i}[e^i - e^{-i}] = 2\pi i \sin(1)\end{aligned}$$

Second Method.

$$\begin{aligned}\int_c \frac{e^z}{z^2+1}dz &= \int_c \frac{e^z d_z}{(z+i)(z-i)} \\ &= \int_{c1} \frac{\frac{e^z}{z-i}}{z+i}dz + \int_c \frac{\frac{e^z}{z+i}}{z-i}dz \\ &= 2\pi i(\frac{e^z}{z-i})_{z=-i} + 2\pi i(\frac{e^z}{z+i})_{z=i} \\ &= [2\pi i\frac{e^{-i}}{-i-i} + 2\pi i\frac{e^i}{i+i}] = \pi[-e^{-i}+e^i] \\ &= \pi(2isin1) = 2\pi i \sin 1\end{aligned}$$

□

EXAMPLE 2.7

Evaluate $\int_c \frac{z-1}{(z+1)^2(z-2)}dz$ where c is $|z-i| = 2$.

Solution: The centre of the circle is at $z = i$ and its radius is 2. Poles are obtained by putting the denominator equal to zero.

$$(z+1)^2(z-2) = 0 \quad \Rightarrow \quad z = -1, -1, 2$$

The integral has two Poles at $z = -1$ (second order) and $z = 2$ (simple pole) of which $z = -1$ is inside the given circle. We can rewrite

$$\int_c \frac{(z-1)dz}{(z+1)^2(z-2)} = \int_{c1} \frac{\frac{z-1}{z-2}}{(z+1)^2} dz$$

By Cauchy Integral formula

$$\int \frac{f(z)}{(z+1)^2} dz = 2\pi i f'(-1)$$

Here

$$f(z) = \frac{z-1}{z-2}$$

$$f'(z) = \frac{(z-2).1-(z-1).1}{(z-2)^2} = \frac{-1}{(z-2)^2} = \frac{-1}{(z-2)^2}$$

$$\Rightarrow \quad f'(-1) = \frac{-1}{(-1-2)^2} = \frac{-1}{9}$$

$$\therefore \int \frac{(z-1)}{(z+1)^2(z-2)} dz == -\frac{2\pi i}{9}$$

□

■ EXAMPLE 2.8

Use Cauchy integral formula to evaluate .

$$\int_c \frac{sin\pi z^2 + cos\pi z^2}{(z-1)(z-2)} dz$$

where c is the circle $|z| = 3$.

Solution: Poles of the integrand are given by putting the denominator equal to zero.

$$(z-1)(z-2) = 0, \quad z = 1, 2$$

The integrand has two poles at $z = 1, 2$. The given circle $|z| = 3$ with centre at $z = 0$ and radius 3 encloses both the poles $z = 1$, and $z = 2$.

$$\int_c \frac{sin\pi z^2 + cos\pi z^2}{(z-1)(z-2)} dz = \int_{c_1} \frac{\frac{sin\pi z^2 + cos\pi z^2}{(z-2)}}{(z-1)} dz + \int_{c_2} \frac{\frac{sin\pi z^2 + cos\pi z^2}{(z-1)}}{(z-2)} dx$$

$$= 2\pi i \left[\frac{sin\pi z^2 + cos\pi z^2}{z-2}\right]_{z=1} + 2\pi i \left[\frac{sin\pi z^2 + cos\pi z^2}{z-1}\right]_{z=2}$$

$$= 2\pi i \left[\frac{sin\pi + cos\pi}{1-2}\right] + 2\pi i \left[\frac{sin4\pi + cos4\pi}{2-1}\right]$$

$$= 2\pi i \left(\frac{-1}{-1}\right) + 2\pi i \left(\frac{1}{1}\right) = 4\pi i$$

□

Problems

2.4 Evaluate the following by using Cauchy integral formula:

a) $\int_c \frac{1}{z-a}dz$, where c is a simple closed curve and the point $z = a$ is (i) outside c; (ii) inside c.

b) $\int_c \frac{e^z}{z-1}dz$, where c is the circle $|z| = 2$.

c) $\int_c \frac{cos\pi z}{z-1}dz$, where c is the circle $|z| = 3$.

d) $\int_c \frac{cos\pi z^2}{(z-1)(z-2)}dz$, where c is the circle $|z| = 3$.

e) $\int_c \frac{e^{-z}}{(z+2)^5}dz$, where c is the circle $|z| = 3$.

f) $\int_c \frac{e^{2z}}{(z+1)^4}dz$, where c is the circle $|z| = 2$.

g) $\int_c \frac{3z^2+z}{z^2-1}dz$, where c is the circle $|z-1| = 1$.

2.5 Evaluate the following integral using Cauchy integral formula

$$\int_c \frac{4-3z}{z(z-1)(z-2)}dz$$

where c is the circle $|z| = \frac{3}{2}$.

2.6 Integrate $\frac{1}{(z^3-1)^2}$ the counter clockwise sense around the circle $|z-1| = 1$.

2.7 Find the value of $\int_c \frac{2z^2+z}{z^2-1}dz$, if c is circle of unit radius with centre at $z = 1$.

CHAPTER 3

SERIES

In this chapter, we discuss the power series expansion of a complex function, viz, the Taylor's and Laurent's series

3.1 TAYLOR'S THEOREM

Theorem 6. *If function $f(z)$ is analytic at all points inside a circle c, with its centre at the point a and radius R, then at each point z inside c,*

$$f(z) = f(a) + (z-a)f'(a) + \frac{(z-a)^2}{2!}f''(a) + \ldots + \frac{(z-a)^n}{n!}f^n(a) + \ldots$$

Proof. Let $f(z)$ be analytic function within and on the circle circle c of radius r centered at a so that

$$c : |z-a| = r$$

Draw another circle $\gamma : |z-a| = \rho$, where $\rho < r$ Hence by Cauchy's integral formula, we have

$$f(z) = \frac{1}{2\pi i}\int_{\gamma}\frac{f(w)}{w-z}dw$$

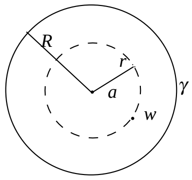


Figure 3.1. Tayer's Theorem

But,

$$\begin{aligned}\frac{1}{w-z} &= \frac{1}{(w-a)-(z-a)}\\ &= \frac{1}{(w-a)}\frac{1}{\left[1-\frac{z-a}{w-a}\right]}\\ &= \frac{1}{(w-a)}\left[1-\frac{z-a}{w-a}\right]^{-1}\\ &= \frac{1}{(w-a)}+\frac{(z-a)}{(w-a)^2}+\frac{(z-a)^2}{(w-a)^3}+\ldots\\ &\qquad +\frac{(z-a)^{n-1}}{(w-a)^n}+\frac{(z-a)^n}{(w-a)^{n+1}}\left[1-\frac{z-a}{w-a}\right]^{-1}\end{aligned}$$

Hence,

$$\frac{1}{2\pi i}\int_\gamma \frac{f(w)}{w-z}dw = \frac{1}{2\pi i}\int_\gamma \frac{f(w)}{(w-a)}dw+\frac{(z-a)}{2\pi i}\int_\gamma \frac{f(w)}{(w-a)^2}dw+\frac{(z-a)^2}{2\pi i}\int_\gamma \frac{f(w)}{(w-a)^3}dw$$
$$+\ldots+\frac{(z-a)^{n-1}}{2\pi i}\int_\gamma \frac{f(w)}{(w-a)^n}dw+R_n$$
$$\text{where } R_n = \frac{(z-a)^n}{2\pi i}\int_\gamma \frac{f(w)}{(w-z)(w-a)^n}dw$$

But, we have

$$\frac{1}{2\pi i}\int_\gamma \frac{f(w)}{(w-z)^{n+1}}dw = \frac{1}{n!}f^{(n)}(z)\text{Hence,}$$

$$f(z) = f(a)+\frac{(z-a)}{1!}f'(a)+\ldots+\frac{(z-a)^{n-1}}{(n-1)!}f^{(n-1)}(a)+R_n$$

which is called Taylor's formula, R_n being the *remainder*.

Now, let M denotes the maximum value of $f(w)$ on γ. Since $|z-a| = \rho$, $|w-a| = r$ and $|w-z| > r-\rho$, hence

$$\begin{aligned}|R_n| &\le \frac{|(z-a)^n|}{|2\pi i|}\int_\gamma \frac{|f(w)|}{|(w-z)||(w-a)^n|}|dw|\\ &\le \frac{|\rho^n|}{2\pi}\frac{M}{(r-\rho)r^n}\int_\gamma |dw|\\ &= \frac{M}{1-\frac{\rho}{r}}\left(\frac{\rho}{r}\right)^n\end{aligned}$$

which tends to zero as n tends to infinity since $\frac{\rho}{r} < 1$. Thus

$$f(z) = f(a) + (z-a)f'(a) + \frac{(z-a)^2}{2!}f''(a) + \ldots + \frac{(z-a)^n}{n!}f^n(a) + \ldots$$

Remark: If $a = 0$, Taylor's series is

$$f(z) = f(0) + (z)f'(0) + \frac{(z)^2}{2!}f''(0) + \ldots + \frac{(z)^n}{n!}f^n(0) + \ldots$$

which is called the *Maclaurin's series* of $f(z)$. □

■ EXAMPLE 3.1

Find the first four terms of the Taylor's series expansion of the complex variable function

$$f(z) = \frac{z+1}{(z-3)(z-4)}$$

about $z = 2$. find the region of convergence.

Solution: Given $f(z) = \dfrac{z+1}{(z-3)(z-4)}$

If centre of a circle is at $z = 2$, then the distance of the singularities $z = 3$ and $z = 4$ from the centre are 1 and 2. Hence if a circle ($|z-2| = 1$) of radius 1 and with centre $z = 2$ is drawn, the given function function $f(z)$ is analytic, hence it can be expanded in a Taylor's series within the circle $|z-2| = 1$, which is therefore the circle of convergence.

$$\begin{aligned}
f(z) &= \frac{z+1}{(z-3)(z-4)} = \frac{-4}{z-3} + \frac{5}{z-4} \quad \text{Using partial fraction} \\
&= \frac{-4}{(z-2)-1} + \frac{5}{(z-2)-2} \\
&= 4[1-(z-2)]^{-1} - \frac{5}{2}\left[1 - \frac{z-2}{2}\right]^{-1} \\
&= 4[1+(z-2)+(z-2)^2+(z-2)^3+\ldots] - \frac{5}{2}\left[1 + \frac{z-2}{2} + \frac{(z-2)^2}{4} + \frac{(z-2)^3}{8} + \ldots\right] \\
&= \frac{3}{2} + \frac{11}{4}(z-2) + \frac{27}{8}(z-2)^2 + \frac{59}{16}(z-2)^3
\end{aligned}$$

□

■ EXAMPLE 3.2

Find the first three terms of the Taylor series expansion of $f(z) = \frac{1}{z^2+4}$ about $z = -i$. Find the region of convergence.

Solution: Here

$$f(z) = \frac{1}{z^2+4}$$

Poles are given by

$$z^2 + 4 = 0 \quad \Rightarrow \quad z = \pm 2i$$

If the centre of a circle is $z = -i$, then the distance of the singularities $z = 2i$ and $z = -2i$ from the centre are 3 and 1. Hence if a circle of radius 1 is drawn at the centre $z = -i$, then within the circle

$|z+i|$=1, the given function $f(z)$ is analytic. Thus the function can be expanded in Taylor Series within the circle $|z+i|=1$, which is therefore the region of convergence ($|z+i|<1$).
This problem could be solved as previous example. Here we use another alternative approach.
We have Taylor Series about $z=-i$,

$$f(z)=f(-i)+(z+i)\frac{f'(-i)}{1!}+(z+i)^2\frac{f''(-i)}{2!}+\ldots$$

Now since $f(z)=\dfrac{1}{z^2+4}$,

$$f(-i)=\frac{1}{3}$$
$$f'(z)=\frac{-2z}{(z^2+4)^2} \quad\Rightarrow\quad f'(-i)=\frac{2i}{9}$$
$$f''(z)=\frac{2(z^2+4)-8z^2}{(z^2+4)^3} \quad\Rightarrow\quad f''(-i)=-\frac{14}{27}$$

Hence Taylor series,

$$f(z)=\frac{1}{3}+\frac{2i}{9}(z+i)+\frac{7}{27}(z+i)^2+\ldots$$

□

3.2 LAURENT'S THEOREM

In expanding a function $f(Z)$ by Taylor's series at a point $z=a$, we require that function $f(z)$ is analytic at $z=a$. Laurent's series gives an expansion of $f(z)$ at a point $z=a$ even if $f(z)$ is not analytic there.

Theorem 7. *If function $f(z)$ is analytic between and on two circles c_1 and c_2 having common centre at $z=a$ and radii r_1 and r_2, then for all points z in this region, $f(z)$ can be expanded by*

$$f(z)=\sum_{n=0}^{\infty}a_n(z-a)^n+\sum_{n=1}^{\infty}\frac{b_n}{(z-a)^n}$$

where

$$a_n=\frac{1}{2\pi i}\int_{c_1}\frac{f(w)}{(w-a)^{n+1}}dw$$
$$b_n=\frac{1}{2\pi i}\int_{c_2}\frac{f(w)}{(w-a)^{-n+1}}dw$$

the integrals being taken in counter clockwise sense.

Proof. Let $f(z)$ be analytic function between and on the circles c_1 and c_2 of radii r_1 and r_2 centered at a so that

$$c_1:|z-a|=r_1 \qquad c_2:|z-a|=r_2$$

Draw another circle $\gamma:|z-a|=\rho$, where $r_2<\rho<r_1$. Let w be a point on circle γ, then by Cauchy's integral formula, we have

$$f(z)=\frac{1}{2\pi i}\int_{c_1}\frac{f(w)}{w-z}dw-\frac{1}{2\pi i}\int_{c_2}\frac{f(w)}{w-z}dw \tag{3.1}$$

But,

$$\begin{aligned}
\frac{1}{w-z} &= \frac{1}{(w-a)-(z-a)} \\
&= \frac{1}{(w-a)} \frac{1}{\left[1-\frac{z-a}{w-a}\right]} \\
&= \frac{1}{(w-a)} \left[1-\frac{z-a}{w-a}\right]^{-1} \\
&= \frac{1}{(w-a)} + \frac{(z-a)}{(w-a)^2} + \frac{(z-a)^2}{(w-a)^3} + \ldots \\
&\qquad + \frac{(z-a)^{n-1}}{(w-a)^n} + \frac{(z-a)^n}{(w-a)^{n+1}} \left[1-\frac{z-a}{w-a}\right]^{-1}
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{1}{2\pi i}\int_{c_1} \frac{f(w)}{w-z}dw &= \frac{1}{2\pi i}\int_{c_1} \frac{f(w)}{(w-a)}dw + \frac{(z-a)}{2\pi i}\int_{c_1} \frac{f(w)}{(w-a)^2}dw + \frac{(z-a)^2}{2\pi i}\int_{c_1} \frac{f(w)}{(w-a)^3}dw \\
&\quad + \ldots + \frac{(z-a)^{n-1}}{2\pi i}\int_{c_1} \frac{f(w)}{(w-a)^n}dw + R_n \\
&\quad \text{where } R_n = \frac{(z-a)^n}{2\pi i}\int_{c_1} \frac{f(w)}{(w-z)(w-a)^n}dw
\end{aligned}$$

R_n being the *remainder*.

Now, let M denotes the maximum value of $f(w)$ on c_1. Since $|z-a| = \rho$, $|w-a| = r_1$ and $|w-z| > r_1 - \rho$, hence

$$\begin{aligned}
|R_n| &\leq \frac{|(z-a)^n|}{|2\pi i|}\int_{c_1} \frac{|f(w)|}{|(w-z)||(w-a)^n|}|dw| \\
&\leq \frac{|\rho^n|}{2\pi} \frac{M}{(r_1-\rho){r_1}^n}\int_{c_1} |dw| \\
&= \frac{M}{1-\frac{\rho}{r_1}} \left(\frac{\rho}{r_1}\right)^n
\end{aligned}$$

which tends to zero as n tends to infinity since $\dfrac{\rho}{r_1} < 1$. Thus

$$\frac{1}{2\pi i}\int_{c_1} \frac{f(w)}{w-z}dw = a_0 + a_1(z-a) + a_2(z-a)^2 + \ldots + a_n(z-a)^n + \ldots \tag{3.2}$$

where

$$a_n = \frac{1}{2\pi i}\int_{c_1} \frac{f(w)}{(w-a)^{n+1}}dw$$

Similarly, since

$$\begin{aligned}
\frac{1}{w-z} &= \frac{1}{(w-a)-(z-a)} \\
&= -\frac{1}{(z-a)}\frac{1}{\left[1-\frac{w-a}{z-a}\right]} \\
&= -\frac{1}{(z-a)}\left[1-\frac{w-a}{z-a}\right]^{-1} \\
&= -\left[\frac{1}{(z-a)}+\frac{(w-a)}{(z-a)^2}+\frac{(w-a)^2}{(z-a)^3}+...\right]
\end{aligned}$$

Hence,

$$\begin{aligned}
-\frac{1}{2\pi i}\int_{c_2}\frac{f(w)}{w-z}dw &= \frac{1}{2\pi i}\int_{c_2}\frac{f(w)}{(z-a)}dw+\frac{1}{2\pi i}\int_{c_2}\frac{(w-a)f(w)}{(z-a)^2}dw+\frac{1}{2\pi i}\int_{c_2}\frac{(w-a)^2 f(w)}{(z-a)^3}dw \\
&+...+\frac{1}{2\pi i}\int_{c_2}\frac{(w-a)^{n-1}f(w)}{(z-a)^n}dw+R_n' \\
&\text{where } R_n' = \frac{1}{2\pi i}\int_{c_2}\frac{(w-a)^n f(w)}{(w-z)(z-a)^n}dw
\end{aligned}$$

Now, let M denotes the maximum value of $f(w)$ on c_2. Since $|z-a| = \rho$, $|w-a| = r_2$ and $|w-z| > r_2-\rho$, hence

$$\begin{aligned}
|R_n'| &\leq \frac{1}{|2\pi i|}\int_{c_1}\frac{|(w-a)^n||f(w)|}{|(w-z)||(z-a)^n|}|dw| \\
&\leq \frac{|r_2^n|}{2\pi}\frac{M}{(r_2-\rho)\rho^n}\int_{c_1}|dw| \\
&= \frac{M}{1-\frac{\rho}{r_2}}\left(\frac{r}{\rho}\right)^n
\end{aligned}$$

which tends to zero as n tends to infinity since $\dfrac{r_2}{\rho} < 1$. Thus

$$-\frac{1}{2\pi i}\int_{c_1}\frac{f(w)}{w-z}dw = \frac{b_1}{(z-a)}+\frac{b_2}{(z-a)^2}+...+\frac{b_n}{(z-a)^n}+... \tag{3.3}$$

where

$$b_n = \frac{1}{2\pi i}\int_{c_1}\frac{f(w)}{(w-a)^{-n+1}}dw$$

Hence from equation 3.1, 3.2 and 3.3, we get

$$f(z) = a_0+a_1(z-a)+a_2(z-a)^2+...+a_n(z-a)^n+...+\frac{b_1}{z-a}+\frac{b_2}{(z-a)^2}+...+\frac{b_n}{(z-a)^n}+...$$

where $a_n = \dfrac{1}{2\pi i}\displaystyle\int_{c_1}\frac{f(w)}{(w-a)^{n+1}}dw$ and $b_n = \dfrac{1}{2\pi i}\displaystyle\int_{c_2}\frac{f(w)}{(w-a)^{-n+1}}dw.$ □

■ EXAMPLE 3.3

Show that, if $f(z)$ has a pole at $z = a$ then $|f(z)| \to \infty$ as $z \to a$.

Solution: Suppose the pole is of order m, then

$$f(z) = \sum_{n=0}^{\infty} a_n(z-a)^n + \sum_{n=1}^{m} b_n(z-a)^{-n}$$

Its principal part is $\sum_{n=1}^{m} b_n(z-a)^{-n}$

$$\begin{aligned}
\sum_{n=1}^{m} b_n(z-a)^{-n} &= \frac{b_1}{z-a} + \frac{b_2}{(z-a)^2} + \ldots + \frac{b_n}{(z-a)^m} \\
&= \frac{1}{(z-a)^m}[b_m + b_{m-1}(z-a) + \ldots b_1(z-a)^{m-1}] \\
&= \frac{1}{(z-a)^m}[b_m + \sum_{n=1}^{m-1} b_n(z-a)^{m-n}] \\
|\sum_{n=1}^{m} b_n(z-a)^{-n}| &= |\frac{z}{(z-a)^m}[b_m + \sum_{n=1}^{m-1} b_n(z-a)^{m-n}]| \\
&\geq |\frac{1}{(z-a)^m}| \left[|b_m| - \sum_{n-1}^{m-1} |b_n||z-a|^{m-n}\right]
\end{aligned}$$

This tends to $b_m|a_1 + a_2| \geq |a_1| - |a_2|$. As $z \to -a$ R.H.S. $=\infty$ □

■ EXAMPLE 3.4

Write all possible Laurent series for the function

$$f(z) = \frac{1}{z(z+2)^3}$$

about the pole $z = -2$. Using appropriate Laurent series.

Solution: To expand $\frac{1}{z(z+2)^3}$ about $z = -2$, i.e., in powers of $(z+2)$, we put $z+2 = t$. Then

$$\begin{aligned}
f(z) &= \frac{1}{z(z+2)^3} = \frac{1}{(t-2)t^3} = \frac{1}{t^3}.\frac{1}{t-2} \\
&= \frac{1}{t^3}.\frac{1}{-2}.\frac{1}{1-\frac{t}{2}} = -\frac{1}{2t^3}(1-\frac{t}{2})^{-1}
\end{aligned}$$

$0 < |z+2| < 1$ or $0 < |t| < 1$

$$\begin{aligned}
f(z) &= -\frac{1}{2t^3}[1 + \frac{t}{2} + \frac{t^2}{4} + \frac{t^3}{8} + \frac{t^4}{16} + \frac{t^5}{32} + \ldots] \\
&= -\frac{1}{2t^3} - \frac{1}{4t^2} - \frac{1}{8t} - \frac{1}{16} - \frac{t}{32} - \frac{t^2}{64} \ldots \\
&= -\frac{1}{2(z+2)^3} - \frac{1}{4(z+2)^2} - \frac{1}{8(z+2)} - \frac{1}{16} - \frac{z+2}{32} - \frac{(z+2)^2}{64} \ldots
\end{aligned}$$

□

EXAMPLE 3.5

Expand $f(z) = \cosh\left(z + \frac{1}{z}\right)$

Solution: $f(z)$ is analytic except $z = 0$. $f(z)$ can be expanded by Laurent's theorem.

$$f(z) = \sum_{n=0}^{\infty} a_n z^n + \sum_{n=1}^{\infty} \frac{b_n}{z^n}$$

where

$$a_n = \frac{1}{2\pi i}\int_c \frac{f(z)dz}{(z-0)^{n+1}}, \text{ and } b_n = \frac{1}{2\pi i}\int_c f(z)z^{n-1}dz$$

$$\begin{aligned}
a_n &= \frac{1}{2\pi i}\int_c \frac{\cosh(z+\frac{1}{z})dz}{z^{n+1}} \\
&= \frac{1}{2\pi i}\int_0^{2\pi} \frac{\cosh(2\cos\theta)dz}{z^{n+1}} \\
&\qquad\qquad\qquad z = e^{i\theta} \\
&\qquad\qquad\qquad dz = ie^{i\theta}d\theta \\
&= \frac{1}{2\pi i}\int_0^{2\pi} \frac{\cosh(2\cos\theta)e^{i\theta}d\theta}{e^{i(n+1)\theta}} \\
&= \frac{1}{2\pi}\int_0^{2\pi} \cosh(2\cos)e^{-ni\theta}d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} \cosh(2\cos\theta)\cos n\theta d\theta - \frac{i}{2\pi}\int_c \cosh(2\theta)\sin n\theta d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} \cosh(2\cos\theta)\cos n\theta d\theta + 0
\end{aligned}$$

Since,

$$\begin{aligned}
b_n &= a_{-n} \\
&= \frac{1}{2\pi}\int_0^{2\pi} \cosh(2\cos\theta)\cos(-n\theta)d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} \cosh(2\cos\theta)\cos(n\theta)d\theta \\
&= a_n
\end{aligned}$$

Therefore,

$$\begin{aligned}
f(z) &= \sum_{n=0}^{\infty} a_n z^n + \sum_{n=1}^{\infty} \frac{b_n}{z^n} = a_0 + \sum_{n=1}^{\infty} a_n z^n + \sum_{n=1}^{\infty} \frac{a_n}{z^n} \\
&= a_0 + \sum_{n=1}^{\infty} a_n(z^n + z^{-n})
\end{aligned}$$

□

EXAMPLE 3.6

Expand $f(z) = e^{\frac{c}{2}\left(z-\frac{1}{z}\right)} = \sum_{-\infty}^{\infty} a_n z^n.$, where $a_n = \frac{1}{2\pi}\cos(n\theta - c\sin\theta)d\theta.$

Solution: $f(z)$ is the analytic function except at $z = 0$ so $f(z)$ can be expanded by Laurent's series.

$$f(z) = \sum_{n=0}^{\infty} a_n z^n + \sum_{n=1}^{\infty} \frac{b_n}{z^n}$$

where $a_n = \frac{1}{2\pi i}\int_c \frac{f(z)dz}{z^{n+1}}$ and $b_n = \frac{1}{2\pi i}\int_c \frac{f(z)dz}{z^{-n+1}}$

Since $f(z)$ remains unaltered if $\frac{-1}{z}$ is written for z, hence

$$b_n = (-1)^n a^n$$

$$\therefore \quad f(z) = \sum_{n=0}^{\infty} a_n z^n + \sum_{n=0}^{\infty} \frac{b_n}{z^n}$$

$$= \sum_{n=0}^{\infty} a_n z^n + (-1)\sum_{n=0}^{\infty} \frac{a_n}{z^n}$$

$$= \sum_{-\infty}^{\infty} a_n z^n$$

$$\begin{aligned}
a_n &= \frac{1}{2\pi i}\int_c \frac{f(z)dz}{z^{n+1}} \\
&= \frac{1}{2\pi i}\int_c \frac{e^{\frac{c}{2}(z-\frac{1}{z})}dx}{z^{n+1}} \\
&= \frac{1}{2\pi i}\int_0^{2\pi} \frac{e^{\frac{c}{2}(2isin\theta)} ie^{i\theta} d\theta}{e^{(n+1)i\theta}} \\
&= \frac{1}{2\pi}\int_0^{2\pi} e^{cisin\theta} e^{-sini\theta} d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} e^{i(csin\theta)-ni\theta} d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} [cos(csin\theta - n\theta) - isin(csin\theta - n\theta)] d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} cos(csin\theta - n\theta) d\theta
\end{aligned}$$

$$\because \text{ if, } f(2a-x) = -f(x), then \int_a^{2a} f(x)dx = 0$$

□

Problems

3.1 Expand $f(z) = \frac{1}{(z-1)(z-2)}$ for $1 < |z| < 2$.

3.2 Obtain the Taylor's or Laurent's series which represents the function $f(z) = \dfrac{1}{(1+z^2)(z+2)}$ when (i) $1 < |z| < 2$ (ii) $|z| > 2$.

3.3 Expand $\dfrac{e^z}{(z-1)^2}$ about $z = 1$

3.4 Expand $f(z) = \sin\left\{c\left(z + \frac{1}{z}\right)\right\}$

3.5 Expand $f(z) = e^{\frac{c}{2}\left(z - \frac{1}{z}\right)}$

3.6 $\dfrac{z-1}{z+1}$ (a) about $z = 0$ (b) about $z = 1$

3.7 Expand $\frac{1}{(z+1)(z+3)}$ in Laurent's series, if (a) $1 < |z| < 3$ (b) $|z| < 3$ (c) $-3 < z < 3$ (d) $|z| < 1$.

3.8 Find the Taylor's and Laurent's series which represents the function $\frac{z^2-1}{(z+2)(z+3)}$ when (i) $|z| < 2$ (ii) $2|z| < 3$.

3.9 Expand $\frac{z}{(z^2+1)(z^2+4)}$ in $1 < |z| < 2$.

3.10 Represent the function f(z) = $\frac{4z+3}{z(z-3)(z+2)}$ in Laurent's series (i) within $|z| = 1$ (ii) in the annular region between $|z| = 2$ and $|z| = 3$.

3.11 Write all possible Laurent Series for the function f(z) = $\frac{z^2}{(z-1)^2(z+3)}$ about the singularity z=1, stating the region of convergence in each case.

3.12 Obtain the expansion

$$f(z) = f(a) + 2\left\{\frac{z-a}{2} f'\left(\frac{z+a}{2}\right) + \frac{(z-a)^3}{2^3.3} f'''\left(\frac{z+a}{2}\right) + \frac{(z-a)^5}{2^5.5} f^v\left(\frac{z+a}{2}\right) + \ldots.\right\}$$

3.13 Expand $\frac{z^2-6z-1}{(z-1)(z+2)(z-3)}$ in $3 < |z+2| < 5$.

CHAPTER 4

SINGULARITY, ZEROS AND RESIDUE

4.1 DEFINITIONS

4.1.1 Zeros

Definition 1. *The value of z for which analytic function $f(z) = 0$ is called zero of the function $f(z)$.*

Let z_0 be a zero of an analytic function $f(z)$. Since $f(z)$ is analytic at z_0, there exists a neighbourhood of z_0 at which $f(z)$ can be expanded in a Taylor's series. i.e.,

$$f(z) = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \ldots a_n(z - z_0)^n + \ldots$$

where $|z - z_0| < \rho$, where $a_0 = f(z_0)$ and $a_n = \dfrac{f^{(n)}(z_0)}{n!}$.

Since z_0 is zero of $f(z)$, $f(z_0) = 0$ which gives $a_0 = 0$ and $a_1 \neq 0$, then such z_0 is said to be a *simple zero*. If $a_0 = 0$ and $a_1 = 0$ but $a_2 \neq 0$ then z_0 is called double zero. In general, if $a_0 = a_1 = a_2 = \ldots = a_{m-1} = 0$ but $a_m \neq 0$ then z_0 is called zero of order m. Thus zero of order m may be defined as the condition

$$f(z_0) = f'(z_0) = f''(z_0) = \ldots = f^{(m-1)}(z_0) = 0 \text{ and } f^{(m)}(z_0) \neq 0 \tag{4.1}$$

In this case function $f(z)$ may be rewritten as

$$f(z) = (z - z_0)^m g(z) \tag{4.2}$$

where $g(z)$ is analytic and $g(z_0) = a_m$, which is a non-zero quantity.

■ **EXAMPLE 4.1**

The function $f(z) = \dfrac{1}{(z-1)}$ has a simple zero at infinity.

■ **EXAMPLE 4.2**

The function $f(z) = (z-1)^3$ has a zero of order 3 at $z = 1$.

■ **EXAMPLE 4.3**

Find the zero of the function defined by

$$f(z) = \frac{(z-3)}{z^3} \sin \frac{1}{(z-2)}$$

Solution: To find zero, we have $f(z) = 0$

$$\frac{(z-3)}{z^3} \sin \frac{1}{(z-2)} = 0$$

Hence $(z-3) = 0$ or $\sin \dfrac{1}{(z-2)} = 0 \Rightarrow \dfrac{1}{(z-2)} = n\pi$, where $n = 0, \pm 1, \pm 2, \ldots$ It follows that $z = 3$ or $z = 2 + \dfrac{1}{n\pi}$, where $n = 0, \pm 1, \pm 2, \ldots$ □

4.1.2 Singularity

Definition 2. *If a function $f(z)$ is analytic at every point in the neighbourhood of a point z_0 except at z_0 itself, then z_0 is called a singularity or a singular point of the function.*

A singularity of an analytic function is the point z at which function ceases to be analytic.

4.1.3 Types of Singularities

Isolated and Non-isolated Singularities. Let $z = a$ be a singularity of $f(z)$ and if there is no other singularity in neighbourhood of the point $z = a$, then this point $z = a$ is said to be an *isolated singularity* and otherwise it is termed as *non-isolated singularity*. (Note when a sequence of singularities is obtained the limit point of the sequence is isolated singularity.)

■ **EXAMPLE 4.4**

The function $f(z) = \frac{1}{\sin \frac{\pi}{z}}$ is analytic everywhere except those points at which

$$\sin \frac{\pi}{z} = 0$$

$$\Rightarrow \qquad \frac{\pi}{z} = n\pi \qquad \Rightarrow \qquad z = \frac{1}{n} \qquad (n = 1, 2, 3, \ldots)$$

Thus point $z = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, ..., z = 0$ are the singularity of $f(z)$. Since no other singularity of $f(z)$ in neighbourhood of these points (except $z = 0$), these are isolated singularities. But at $z = 0$, there are infinite number of other singularities, where n is very large. Therefore $z = 0$ is non isolated singularity.

EXAMPLE 4.5

The function $f(z) = \dfrac{1}{(z-a)(z-b)}$ is analytic everywhere except $z = a$ and $z = b$. Thus point $z = a$ and $z = b$ are singularity of $f(z)$. Also there is no other singularity of $f(z)$ in neighbourhood of these points, these are isolated singularities.

Principal Part of $f(z)$ at isolated singularity. Let $f(z)$ be analytic function within domain D except at the point $z = a$ which is an isolated singularity. Now draw a circle C centered at $z = a$ and of radius as small as we please. draw another concentric circle of any radius say R lying wholly within the domain D. The function $f(z)$ is analytic in ring shaped region between these two circles. Hence by Laurent's Theorem, we have

$$f(z) = \sum_{n=0}^{\infty} a_n (z-a)^n + \sum_{n=1}^{\infty} b_n (z-a)^{-n}$$

The second term $\sum_{n=1}^{\infty} b_n (z-a)^{-n}$ in this expansion is called *principal part* of $f(z)$ at the singularity $z = a$.

- If there is no of term in principal part. Then singularity $z = a$ is called *removal singularity*.

$$f(z) = \frac{\sin(z-a)}{(z-a)} = \frac{1}{(z-a)} \left[(z-a) - \frac{(z-a)^3}{3!} + \frac{(z-a)^5}{5!} - ...\infty \right]$$

$$= 1 - \frac{(z-a)^2}{3!} + \frac{(z-a)^4}{5!} - ...\infty$$

- If there are infinite number of terms in principal part. Then singularity $z = a$ is called *essential singularity*.

$$f(z) = e^{\frac{1}{z}} = 1 + \frac{1}{z} + \frac{1}{2!} \frac{1}{z^2} + \frac{1}{3!} \frac{1}{z^3} ...\infty$$

- If there are finite number of term in principal part. (Say m terms). Then singularity $z = a$ is called *pole*. and m is called ORDER of pole.

$$f(z) = \frac{\sin(z-a)}{(z-a)^3} = \frac{1}{(z-a)^3} \left[(z-a) - \frac{(z-a)^3}{3!} + \frac{(z-a)^5}{5!} - ...\infty \right]$$

$$= (z-a)^{-2} - \frac{1}{3!} + \frac{(z-a)^2}{5!} - ... + \infty$$

Definition 3. *A functions is said to be meromorphic function if it has poles as its only type of singularity.*

Definition 4. *A functions is said to be entire function if it has no singularity.*

Definition 5. *Limit point of zeros is called isolated essential singularity.*

Definition 6. *Limit point of poles is called non-isolated essential singularity.*

■ **EXAMPLE 4.6**

Show that the function e^z has an isolated essential singularity at $z = \infty$

Solution: **Note:** The behaviour of function $f(z)$ at ∞ is same as the behaviour $f\left(\frac{1}{z}\right)$ at $z = 0$.
Here

$$f\left(\frac{1}{z}\right) = e^{\frac{1}{z}}$$
$$= \sum_{n=0}^{\infty} \frac{z^{-n}}{n!}$$

Thus, $f\left(\frac{1}{z}\right)$ contains an infinite number of terms in the negative power of z. Therefore, by definition, $z = 0$ is an essential singularity of function $f\left(\frac{1}{z}\right)$. Consequently $f(z)$ has essential singularity at $z = \infty$. □

Problems

4.1 Find out the zeros and discuss the nature of singularities of $f(z) = \frac{z-2}{z^2} \sin \frac{1}{z-1}$

4.2 What kind of singularities, the following functions have:

a) $f(z) = \frac{1}{1-e^z}$ at $z = 2\pi i$

b) $f(z) = \frac{1}{\sin z - \cos z}$ at $z = \frac{\pi}{4}$

c) $f(z) = \cos z - \sin z$ at $z = \infty$

d) $f(z) = \frac{\cot \pi z}{(z-a)^2}$ at $z = 0$ and $z = \infty$

e) $f(z) = \sin \frac{1}{1-z}$ at $z = 1$

f) $f(z) = \tan \frac{1}{z}$ at $z = 0$

g) $f(z) = \frac{1}{\cos \frac{1}{z}}$ at $z = 0$

h) $f(z) = \frac{1}{\sin \frac{1}{z}}$ at $z = 0$

i) $f(z) = \frac{1-e^z}{1+e^z}$ at $z = \infty$

j) $f(z) = z$ cosecz at $z = \infty$

4.2 THE RESIDUE AT POLES

The coefficient of $\frac{1}{(z-a)}$ in the principal part of Laurent's Expansion is called the RESIDUE of function $f(z)$. The coefficient b_1 which is given as

$$b_1 = \frac{1}{2\pi i} \int_C f(z) dz \quad = RES[f(z)]_{z=a}$$

4.2.1 Methods of finding residue at poles

4.2.1.1 The residue at a **simple pole** *(Pole of order one.)*

$$R = \lim_{z \to a} [(z-a).f(z)]$$

If function of form $f(z) = \frac{\phi(z)}{\psi(z)}$ Then

$$R = \left[\frac{\phi(z)}{\psi'(z)}\right]_{z=a}$$

4.2.1.2 The residue at multiple pole (Pole of order m.)

$$R = \frac{1}{(m-1)!}\left[\frac{d^{m-1}}{dz^{m-1}}(z-a)^m.f(z)\right]_{z=a}$$

4.2.1.3 The residue at pole of any order Pole is $z = a$

Put $z - a = t$. Expand function. Now the coefficient of $1/t$ is residue.

EXAMPLE 4.7

Determine the poles and the residue at each pole of the function

$$f(z) = \frac{z^2}{(z-1)^2(z+2)}$$

Solution: The poles of the function $f(z)$ are given by putting the denominator equal to zero. i.e.,

$$(z-1)^2(z+2) = 0$$

$$z = 1, 1, -2$$

The function $f(z)$ has a simple pole[1] at $z = -2$ and pole of order 2 at $z = 1$.

$$\begin{aligned}
\text{Residue of } f(z) \text{ at } (z=-2) &= \lim_{z \to -2}(z+2).f(z) \\
&= \lim_{z \to -2}(z+2)\frac{z^2}{(z-1)^2(z+2)} \\
&= \lim_{z \to -2}\frac{z^2}{(z-1)^2} = \frac{4}{9}
\end{aligned}$$

$$\begin{aligned}
\text{Residue of } f(z) \text{ at double pole (order 2) } (z=1) &= \frac{1}{(2-1)!}\left[\frac{d^{2-1}}{dz^{2-1}}(z-1)^2.f(z)\right]_{z=1} \\
&= \left[\frac{d}{dz}(z-1)^2\frac{z^2}{(z-1)^2(z+2)}\right]_{z=1} \\
&= \left[\frac{d}{dz}\frac{z^2}{(z+2)}\right]_{z=1} \\
&= \left[\frac{z^2+4z}{(z+2)^2}\right]_{z=1} = \frac{5}{9}
\end{aligned}$$

□

[1] Pole of order one is called simple pole.

■ EXAMPLE 4.8

Determine the poles and residue at each pole of the function $f(z) = \cot z$

Solution: Here $\cot z = \dfrac{\cos z}{\sin z}$. i.e. $\phi(z) = \cos z$ and $\psi(z) = \sin z$. The poles of the function $f(z)$ are given by

$$\sin z = 0 \Rightarrow z = n\pi, \text{ where } n = 0, \pm 1, \pm 2, \ldots$$

$$\begin{aligned}
\text{Residue of } f(z) \text{ at } (z = n\pi) &= \left[\frac{\phi(z)}{\psi'(z)}\right]_{z=n\pi} \\
&= \left[\frac{\cos z}{\frac{d}{dz}\sin z}\right]_{z=n\pi} \\
&= \frac{\cos z}{\cos z} = 1
\end{aligned}$$

□

■ EXAMPLE 4.9

Find the residue of $\dfrac{ze^z}{(z-a)^3}$ at its poles.

Solution: The pole of $f(z)$ is given by $(z-a)^3 = 0$, i.e., $z = a$ (pole of order 3)
Putting $z = t + a$

$$\begin{aligned}
f(z) &= \frac{ze^z}{(z-a)^3} \\
\Rightarrow f(z) &= \frac{(t+a)e^{t+a}}{t^3} \\
&= \left(\frac{a}{t^3} + \frac{1}{t^2}\right) e^{t+a} \\
&= e^a \left(\frac{a}{t^3} + \frac{1}{t^2}\right) e^t \\
&= e^a \left(\frac{a}{t^3} + \frac{1}{t^2}\right) \left(1 + \frac{t}{1!} + \frac{t^2}{2!} + \ldots\right) \\
&= e^a \left(\frac{a}{t^3} + \frac{a}{t^2} + \frac{a}{2t} + \frac{1}{t^2} + \frac{1}{t} + \frac{1}{2} + \ldots\right)
\end{aligned}$$

Hence the residue at $(z = a)$ =Coefficient of $\dfrac{1}{t} = e^a \left(\dfrac{a}{2} + 1\right)$. □

4.2.2 Residue at Infinity

Residue of $f(z)$ at $z = \infty$ is defined as $-\dfrac{1}{2\pi i}\int_C f(z)dz$, where integration is taken round in anti-clockwise direction. Where C is a large circle containing all finite singularities of $f(z)$. We also have, Residue at Infinity = $\lim_{z} \rightarrow \infty(-zf(z))$

Residue at Infinity = - Residue at zero of $f(1/z)$

■ **EXAMPLE 4.10**

Find the residue of $f(z) = \dfrac{z^3}{z^2-1}$ at $z = \infty$.

Solution:

Here,

$$f(z) = \frac{z^3}{z^2-1}$$

Therefore

$$f(1/z) = \frac{1}{z(z^2-1)}$$

Residue at $z = 0$

$$\lim_{z \to 0} z . \frac{1}{z(z^2-1)} = -1$$

Hence Residue at Infinity of $f(z)$ = - Residue at zero of $f(1/z)$ = 1 □

Problems

4.3 Determine poles of the following functions. Also find the residue at its poles.

a) $\frac{z-3}{(z-2)^2(z+1)}$

b) $\frac{ze^{iz}}{z^2+a^2}$

c) $\frac{e^z}{(z^2+a^2)}$

d) $\frac{z^2}{(z+2)(z^2+1)}$

e) $z^2 e^{1/z}$

f) $z^2 \sin \frac{1}{z}$

g) $\frac{e^{2z}}{(1+e^z)}$

h) $\frac{1+e^z}{\sin z + z\cos z}$ at $z = 0$

i) $\frac{1}{z(e^z-1)}$

j) $\frac{\cot \pi z}{(z-a)^2}$

4.4 Find the residue at $z = 1$ of $\frac{z^3}{(z-1)^4(z-2)(z-3)}$

4.5 Locate the poles of $\dfrac{e^{az}}{\cosh \pi z}$ and evaluate the residue at the pole of the smallest positive value of z.

4.6 Find the residue of $\frac{z^3}{z^2-1}$ at $z = \infty$

4.7 Find the residue of $\frac{z^2}{(z-a)(z-b)(z-c)}$ at $z = \infty$

4.8 The function $f(z)$ has a double pole at $z = 0$ with residue 2, a simple pole at $z = 1$ with residue 2, is analytic at all finite points of the plane and is bounded as $|z| \to \infty$ if $f(2) = 5$ and $f(-1) = 2$ find $f(z)$

4.9 Let $\dfrac{P(z)}{Q(z)}$, where both $P(z)$ and $Q(x)$ are complex polynomial of degree 2. If $f(0) = f(-1) = 0$. and only singularly of $f(z)$ is of order 2 at $z = 1$ with residue -1, then find $f(z)$.

CHAPTER 5

RESIDUE: EVALUATION OF REAL INTEGRALS

5.1 RESIDUE THEOREM

5.1.1 Residue

By Laurent series expansion of an analytic function $f(z)$, we have

$$f(x) = \sum_{n=-\infty}^{\infty} a_n(z-a)^n$$

where

$$a_n = \frac{1}{2\pi i}\int_C \frac{f(t)}{(t-a)^{n+1}}dt$$

As we have from definition, the coefficient of $\frac{1}{(z-a)}$ in Laurent expansion about $z = a$ is called as RESIDUE, Which may be obtained by putting $n = -1$ in above equation. Thus,

$$RES(a) = \frac{1}{2\pi i}\int_C f(t)dt$$

which implies

$$\int_C f(t)dt = 2\pi i.RES(a)$$

5.1.2 Residue Theorem

Theorem 8. *If $f(z)$ be analytic within and on a simple closed curve C except at number of poles. (Say $z_1, z_2, z_3, \ldots, z_n$ are poles.) then the integral*

$$\int_C f(t)dt = 2\pi i(\text{Sum of Residue of } f(z) \text{ at each poles})$$

Proof. Let $f(z)$ be analytic within and on simple closed curve C except at number of poles $z_1, z_2, z_3, \ldots, z_n$.

Let $C_1, C_2, \ldots, C_n$ be small circles with centres at $z_1, z_2, z_3, \ldots, z_n$ respectively. Then by Cauchy extension theorem, we have

$$\int_C f(z)dz = \int_{C_1} f(z)dz + \int_{C_2} f(z)dz + \ldots + \int_{C_n} f(z)dz$$

Now, we have,

$$\int_{C_1} f(z)dt = 2\pi i.RES(z_1)$$

$$\int_{C_2} f(z)dt = 2\pi i.RES(z_2)$$

$$\ldots \ \ldots \ \ldots$$

$$\int_{C_n} f(z)dt = 2\pi i.RES(z_n)$$

Therefore,

$$\int_C f(z)dz = 2\pi i.RES(z_1) + 2\pi i.RES(z_2) + \ldots + 2\pi i.RES(z_n)$$

$$\int_C f(z)dz = 2\pi i.[RES(z_1) + RES(z_2) + \ldots + RES(z_n)]$$

$$\int_C f(z)dz = 2\pi i.[\text{Sum of all Residues at each poles.}]$$

$$\int_C f(z)dz = 2\pi i.Res$$

We use *Res* for 'Sum of Residues' through out the text. □

5.2 EVALUATION OF REAL INTEGRALS BY CONTOUR INTEGRATION

Type $\int_0^{2\pi} f(\cos\theta, \sin\theta)d\theta$

For such questions consider a unit radius circle with center at origin, as contour

$$|z| = 1 \qquad \Rightarrow z = e^{i\theta} \qquad d\theta = \frac{dz}{iz}$$

Now use following relations

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} = \frac{1}{2}(z + \frac{1}{z})$$

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} = \frac{1}{2i}(z - \frac{1}{z})$$

We get, the whole function is converted into a function of $f(z)$, Now integral become

$$I = \int_C f(z)dz$$

where C is unit circle. The value of this integral may be obtained by using Residue Theorem, which is $2\pi i$.(Sum of Residue inside C)

5.2.1 Form I

Form $\int_0^{2\pi} \frac{1}{a+b\cos\theta} d\theta$ or $\int_0^{2\pi} \frac{1}{a+b\sin\theta} d\theta$

EXAMPLE 5.1

Use residue calculus to evaluate the following integral $\int_0^{2\pi} \frac{1}{5-4sin\theta} d\theta$

Solution: Put $z = e^{i\theta}$ so that $\sin\theta = \frac{1}{2i}\left(z - \frac{1}{z}\right)$ and $d\theta = \frac{dz}{iz}$. Then

$$\int_0^{2\pi} \frac{1}{5-4\sin\theta} d\theta = \oint_c \frac{1}{5-4\left[\frac{1}{2i}\left(z-\frac{1}{z}\right)\right]} \frac{dz}{iz}$$
$$= \oint_c \frac{1}{5iz - 2z^2 + 2} dz$$

Poles of integrand are given by

$$-2z^2 + 5iz + 2 = 0$$

or

$$z = \frac{-5i \pm \sqrt{-25+16}}{-4} = \frac{-5 \pm 3i}{-4} = 2i, \frac{i}{2}$$

Only $z = \frac{i}{2}$ lies inside c. Residue at the simple pole at $z = \frac{i}{2}$ is

$$\lim_{z \to \frac{i}{2}} \left(z - \frac{i}{2}\right) \times \left(\frac{1}{(2z-i)(-z+2i)}\right) = \frac{1}{3i}$$

Hence by Cauchy's residue theorem

$$\int_0^{2\pi} \frac{1}{5-4\sin\theta} d\theta = 2\pi i \times Res$$
$$= 2\pi \times \frac{1}{3i} = \frac{2\pi}{3}$$

□

EXAMPLE 5.2

Using complex variables, evaluate the real integral

$$\int_0^{2\pi} \frac{d\theta}{1 - 2p\sin\theta + p^2}$$

where $p^2 < 1$

Solution:

We have,

$$I = \int_0^{2\pi} \frac{d\theta}{1 - 2p\sin\theta + p^2} = \int_0^{2\pi} \frac{d\theta}{1 - 2p\frac{(e^{i\theta} - e^{-i\theta})}{2i} + p^2}$$

Put $z = e^{i\theta}, dz = ie^{i\theta}d\theta = \frac{dz}{zi}$

$$\begin{aligned} I &= \int_c \frac{1}{1 + ip(z - \frac{1}{z}) + p^2} \frac{dz}{zi} \\ &= \int_c \frac{dz}{zi - pz^2 + p + p^2 zi} = \int_c \frac{dz}{(iz + p)(izp + 1)} \end{aligned}$$

Pole are given by

$$\begin{aligned} (iz + p)(ipz + 1) &= 0 \\ z = -\frac{p}{i}, -\frac{1}{ip} &\text{ or } z = ip, \frac{i}{p} \end{aligned}$$

ip is the only poles inside the unit circle. Residue at$(z = ip)$

$$= \lim_{z \to ip} \frac{(z - pi)}{(iz + p)(izp + 1)} = \lim_{z \to ip} \frac{1}{i(izp + 1)} = \frac{1}{i}\left(\frac{1}{-p^2 + 1}\right)$$

Hence by Cauchy's residue theorem

$$\int_0^{2\pi} \frac{d\theta}{1 - 2psin + p^2} = 2\pi i \times Res = 2\pi i \times \frac{1}{i}\left(\frac{1}{1 - p^2}\right) = \frac{2\pi}{1 - p^2}$$

□

5.2.2 Form II

Form $\int_0^{2\pi} \frac{\cos m\theta}{a + b\cos\theta}d\theta$, $\int_0^{2\pi} \frac{\cos m\theta}{a + b\sin\theta}d\theta$, $\int_0^{2\pi} \frac{\sin m\theta}{a + b\cos\theta}d\theta$ or $\int_0^{2\pi} \frac{\sin m\theta}{a + b\sin\theta}d\theta$

■ **EXAMPLE 5.3**

Using complex variable techniques evaluate the real integral $\int_0^{2\pi} \frac{\sin^2\theta d\theta}{5 - 4\cos\theta}$

Solution: We may write,

$$\int_0^{2\pi} \frac{\sin^2\theta d\theta}{5 - 4\cos\theta} = \frac{1}{2}\int_0^{2\pi} \frac{1 - \cos 2\theta}{5 - 4\cos\theta}d\theta$$

Put $z = e^{i\theta}$so that $\sin\theta = \frac{1}{2i}\left(z - \frac{1}{z}\right)$and $d\theta = \frac{dz}{iz}$. Then

$$\begin{aligned} \frac{1}{2}\int_0^{2\pi} \frac{1 - \cos 2\theta}{5 - 4\cos\theta}d\theta &= \text{Real part of } \frac{1}{2}\int_0^{2\pi} \frac{1 - e^{2i\theta}}{5 - 4cos\theta}d\theta \\ &= \text{Real part of } \frac{1}{2}\oint_c \frac{1 - z^2}{5 - 2(z + \frac{1}{z})}\left(\frac{dz}{iz}\right) \\ &= \text{Real part of } \frac{1}{2i}\oint_c \frac{z^2 - 1}{2z^2 - 5z + 2}dz \end{aligned}$$

Poles of integrand are given by

$$2z^2 - 5z + 2 = 0$$
$$(2z-1)(z-2) = 0$$
$$z = \frac{1}{2}, 2$$

So poles inside the contour c there is a simple pole at $z = \frac{1}{2}$. Residue at the simple pole $(z = \frac{1}{2})$ is

$$\lim_{z\to\frac{1}{2}} (z - \frac{1}{2}) \frac{z^2-1}{(2z-1)(z-2)}$$
$$= \lim_{z\to\frac{1}{2}} \frac{z^2-1}{2(z-2)} = \frac{\frac{1}{4}-1}{2(\frac{1}{2}-2)} = \frac{1}{4}$$
$$\oint_c \frac{z^2-1}{2z^2-5z+2} dz = 2\pi i \times Res = \frac{2\pi i}{4}$$
$$\therefore \int_0^{2\pi} \frac{sin^2\theta}{5-4cos\theta} d\theta = \frac{1}{2i} \oint_c \frac{z^2-1}{2z^2-5z+2} dz = \frac{1}{2i} \times \frac{2\pi i}{4} = \frac{\pi}{4}$$

□

5.2.3 Form III

$\int_0^{2\pi} \frac{1}{(a+b\cos\theta)^2} d\theta$ or $\int_0^{2\pi} \frac{1}{(a+b\sin\theta)^2} d\theta$

■ **EXAMPLE 5.4**

Use the residue theorem to show that

$$\int_0^{2\pi} \frac{d\theta}{(a+bcos\theta)^2} = \frac{2\pi a}{(a^2-b^2)^{3/2}}$$

Solution: Put $e^{i\theta} = z$,so that $e^{i\theta}(id\theta) = dz$or $izd\theta = dz = d\theta = \frac{dz}{iz}$

$$\int_0^{2\pi} \frac{d\theta}{(a+bcos\theta)^2} = \int_c \frac{1}{[a+\frac{b}{2}(z+\frac{1}{z})]^2} \frac{dz}{iz}$$

where c is the unit circle $z = 1$

$$= \int_0 \frac{-4izdz}{(bz^2+2az+b)^2} = -\frac{4i}{b^2} \int_c \frac{zdz}{(z^2+\frac{2az}{b}+1)^2}$$

The pole are given by putting the denominator

$$(z^2 + \frac{2a}{b} z + 1)^2 = 0$$

$$(z-\alpha)^2 (z-\beta)^2 = 0$$

where

$$[\alpha + \beta = -\frac{2a}{b}], \alpha\beta = 1$$

$$\alpha = \frac{-\frac{2a}{b} + \sqrt{\frac{4a^2}{b^2} - 4}}{2} = \frac{-a + \sqrt{a^2-b^2}}{b}$$

$$\beta = \frac{-\frac{2a}{b} - \sqrt{\frac{4a^2}{b^2} - 4}}{2} = \frac{-a - \sqrt{a^2 - b^2}}{b}$$

There are two poles at $z = \alpha$ and $z = \beta$, each of order 2.

Now,

Let

$$f(Z) = \frac{-4iz}{b^2(z-\alpha)^2(z-\beta)^2} = \frac{\phi(z)}{(z-\alpha)^2}$$

where $\phi(z) = \frac{-4iz}{b^2(z-\beta)^2}$. Here $z = \alpha$ is only point inside circle c. Residue at the double pole $z = \alpha$

$$= \lim_{z\to\alpha}\left(\frac{d}{dz}(z-\alpha)^2\frac{\phi(z)}{(z-\alpha)^2}\right) = \lim_{z\to\alpha}\frac{d}{dz}\phi(z)$$

$$= -\lim_{z\to\alpha}\frac{4i}{b^2}\left[\frac{(z-\beta)^2.1 - z.2(z-\beta)}{(z-\beta)^4}\right] = -\frac{4i}{b^2}\lim_{z\to\alpha}\left[\frac{(z-\beta)-2z}{(z-\beta)^3}\right]$$

$$= \frac{4i}{b^2}\frac{\alpha+\beta}{(\alpha-\beta)^3} = \frac{4i}{b^2}\frac{-\frac{2a}{b}}{\left(\frac{2\sqrt{a^2-b^2}}{b}\right)^3} = \frac{-ai}{(a^2-b^2)^{\frac{3}{2}}}$$

Hence

$$\int_0^{2\pi}\frac{d\theta}{(a+bcos\theta)^2} = 2\pi i \times \frac{-ai}{(a^2-b^2)^{\frac{3}{2}}} = \frac{2\pi a}{(a^2-b^2)^{\frac{3}{2}}}$$

□

5.2.4 Form IV

$\int_0^{2\pi}\frac{1}{a^2+\cos^2\theta}d\theta$, $\int_0^{2\pi}\frac{1}{a^2+\sin^2\theta}d\theta$

■ **EXAMPLE 5.5**

Show by the method of residue, that

$$\int_0^{\pi}\frac{ad\theta}{a^2+sin^2\theta} = \frac{\pi}{\sqrt{1+a^2}}$$

$$I = \int_0^{\pi}\frac{ad\theta}{a^2+sin^2\theta} = \int_0^{\pi}\frac{2ad\theta}{2a^2+2sin^2\theta}$$

$$\because (cos2\theta = 1 - 2sin^2)$$

$$= \int_0^{\pi}\frac{2ad\theta}{2a^2+1-cos2\theta} = \int_0^{2\pi}\frac{ad\phi}{2a^2+1-cos\phi}$$

Putting $2\theta = \phi$

$$= \int_0^{2\pi}\frac{ad\phi}{2a^2+1-\frac{1}{2}(e^{i\phi}+e^{-i\phi})} = \int_0^{2\pi}\frac{2ad\phi}{4a^2+2-(e^{i\phi}+e^{-i\phi})}$$

Now let $z = e^{i\phi}$ $\Rightarrow d\phi\frac{dz}{iz}$

$$I = \int_0^{2\pi} \frac{2a}{4a^2+2-(z+\frac{1}{z})} \frac{dz}{iz}$$
$$= 2ai \int_0^{2\pi} \frac{dz}{[z^2-(4a^2+2)z-1]}$$

The pole are given by putting the denominator

$$z^2-(4a^2+2)z-1=0$$

$$z = \frac{(4a^2+2) \pm \sqrt{(4a^2+2)^2-4}}{2} = \frac{(4a^2+2) \pm \sqrt{16a^4+16a^2}}{2}$$
$$= (2a^2+1) \pm 2a\sqrt{a^2+1}$$

Let

$$\alpha = (2a^2+1) + 2a\sqrt{a^2+1}$$

$$\beta = (2a^2+1) - 2a\sqrt{a^2+1}$$

Let $z^2-(4a^2+2)z+1 = (z-\alpha)(z-\beta)$. Therefore product of the roots $= \alpha\beta = 1$ or $|\alpha\beta| = 1$. But $|\alpha| > 1$ therefore $|\beta| < 1$.

i.e., Only β lies inside the circle c.

Residue (at $z = \beta$) is

$$= \lim_{z\to\beta}(z-\beta)\frac{2ai}{(z-\alpha)(z-\beta)} = \frac{-2ai}{(\beta-\alpha)}$$
$$= \frac{2ai}{\left[(2a^2+1)-2a\sqrt{a^2+1}\right]-\left[(2a^2+1)+2a\sqrt{a^2+1}\right]}$$
$$= \frac{2ai}{-4a\sqrt{a^2+1}} = \frac{-i}{2\sqrt{a^2+1}}$$

Hence by Cauchy's residue theorem

$$I = 2\pi i \times Res$$

$$= 2\pi i \frac{-i}{2\sqrt{a^2+1}} = \frac{\pi}{\sqrt{a^2+1}}$$

Problems

5.1 Use Residue of calculus to evaluate the following integrals:

a) $\int_0^{\pi} \frac{1}{3+2\cos\theta} d\theta$

b) $\displaystyle\int_0^{2\pi} \frac{\cos\theta}{3+\sin\theta}d\theta$

c) $\displaystyle\int_0^{\pi} \frac{1+2\cos\theta}{5+4\sin\theta}d\theta$

d) $\displaystyle\int_0^{2\pi} \frac{\sin^2\theta-2\cos\theta}{2+\cos\theta}d\theta$

e) $\displaystyle\int_0^{2\pi} \frac{\cos 3\theta}{5-4\cos\theta}d\theta$

f) $\displaystyle\int_0^{2\pi} e^{\cos\theta}[\cos(\sin\theta-n\theta)]d\theta$

g) $\displaystyle\int_0^{2\pi} \frac{(1+2\cos\theta)^n\cos n\theta}{3+2\cos\theta}d\theta$

h) $\displaystyle\int_0^{2\pi} \frac{\cos 2\theta}{1-2p\sin\theta+p^2}d\theta$ where $p^2<1$.

5.2.5 Some Important Results

Jordan's Inequality

Consider the relation $y=\cos\theta$. As θ increases, $\cos\theta$ decreases and therefore y decreases. The mean ordinate between 0 and θ is $\displaystyle\frac{1}{\theta}\int_0^{\theta}\cos\theta d\theta=\frac{\sin\theta}{\theta}$ which implies when

$$0<\theta<\frac{\pi}{2} \text{ then } \frac{2}{\pi}<\frac{\sin\theta}{\theta}<1$$

Theorem 9. *Let AB be the arc $\alpha<\theta<\beta$ of the circle $|z-a|=r$. If $\lim_{z\to a}(z-a)f(z)=k$ then*

$$\lim_{r\to 0}\int_{AB} f(z)dz=i(\beta-\alpha)k$$

Theorem 10. *Let AB be the arc $\alpha<\theta<\beta$ of the circle $|z|=R$. If $\lim_{z\to\infty} z.f(z)=k$ then*

$$\lim_{R\to\infty}\int_{AB} f(z)dz=i(\beta-\alpha)k$$

Theorem 11. *If $f(z)\to 0$ uniformly as $|z|\to\infty$, then $\lim_{R\to\infty}\int_{C_R} e^{imz}f(z)dz=0$, where C_R denotes the semicircle $|z|=R, \quad m>0$.*

5.2.6 Type $\int_{-\infty}^{\infty} f(x)dy$

Consider the integral $\int_{-\infty}^{\infty} f(x)dy$, such that $f(x)=\frac{\phi(x)}{\psi(x)}$, where $\psi(x)$ has no real roots and the degree of $\psi(x)$ is greater than $\phi(x)$.

Procedure. Consider a function $f(z)$ corresponding to function $f(x)$. Again consider the integral $\int_C f(z)dz$, where C is a curve, consisting of upper half of the circle $|z|=R$, and part of real axis from $-R$ to R. Here R is on our choice and can be taken as such that there is no singularity on its circumference C_R.

Now by Cauchy's theorem,we have

$$\int_C f(z)dz=2\pi i.\sum R_k$$

$$\int_{-R}^{R} f(x)dx + \int_{C_R} f(z)dz = 2\pi i.\sum R_k$$

Now as,

$$\lim_{R\to\infty}\int_{C_R} f(z)dz = 0$$

We get,

$$\int_{-\infty}^{\infty} f(x)dx = 2\pi i.\sum R_k$$

Which is required integral.

■ EXAMPLE 5.6

Evaluate $\int_{-\infty}^{\infty}\frac{x^2dx}{(x^2+1)(x^2+4)}$.

Solution: Consider the integral $\int_C f(z)dz$ where $f(z) = \frac{z^2}{(z^2+1)(z^2+4)}$ and C is the contour consisting of the semi circle C_R which is upper half of a large circle $|z| = R$ of radius R together with the part of the real axis from$-R$ to $+R$.

For the poles

$$(z^2+1)(z^2+4) = 0 \;\Rightarrow\; z = \pm i, z = \pm 2i$$

So $z = i, 2i$ are the only poles inside C.

The residue at $z = i$

$$\begin{aligned}&= \lim_{z\to i}(z-i)\frac{z^2}{(z+i)(z-i)(z^2+4)}\\ &= \lim_{z\to i}\frac{z^2}{(z+i)(z^2+4)} = \frac{-1}{2i(-1+4)} = -\frac{1}{6i}\end{aligned}$$

The residue at $z = 2i$

$$\begin{aligned}&= \lim_{z\to 2i}(z-2i).\frac{z^2}{(z^2+1)(z+2i)(z-2i)}\\ &= \lim_{z\to 2i}\frac{z^2}{(z^2+1)(z+2i)} = \frac{(2i)^2}{(-4+1)(2i+2i)} = \frac{1}{3i}\end{aligned}$$

By theorem of residue;

$$\begin{aligned}\int_C f(z)dz &= 2\pi i[Res(i) + Res(2i)]\\ &= 2\pi i(-\frac{1}{6i} + \frac{1}{3i}) = \frac{\pi}{3}\end{aligned}$$

i.e.

$$\int_{-R}^{R} f(x)dx + \int_{C_R} f(z)dz = \frac{\pi}{3}$$

Hence by making $R \to \infty$, relation (1) becomes

$$\int_{-\infty}^{\infty} f(x)dx + \lim_{R\to\infty}\int_{C_R} f(z)dz = \frac{\pi}{3}$$

Now

$$\begin{aligned}
\left|\int_{C_R} f(z)dz\right| &= \left|\int_{C_R} \frac{z^2dz}{(z^2+1)(z^2+4)}\right| \\
&\leq \int_{C_R} \frac{|z^2dz|}{|(z^2+1)|\,|(z^2+4)|} \\
&\leq \int_{C_R} \frac{|z^2dz|}{|(|z|^2-1)|\,|(|z|^2-4)|} \\
&\qquad \because |z| = R \\
&\qquad z = Re^{i\theta}, 0 < \theta < \pi \\
&\qquad |dz| = Rd\theta, \\
&\leq \int_{C_R} \frac{R^2Rd\theta}{(R^2-1)(R^2-4)} \\
&= \frac{\pi R^3}{(R^2-1)(R^2-4)} \to 0 \text{ as } R \to \infty
\end{aligned}$$

Thus

$$\int_{-\infty}^{\infty} f(x)dx = \frac{\pi}{3}$$

i.e.,

$$\int_{-\infty}^{\infty} \frac{x^2dx}{(x^2+1)(x^2+4)} = \frac{\pi}{3}$$

□

■ EXAMPLE 5.7

Evaluate by the model of complex variables , the integral $\displaystyle\int_{-\infty}^{\infty} \frac{x^2}{(1+x^2)^3}dx$

Solution: Consider the integral $\int_C f(z)dz$ where $f(z) = \frac{z^2}{(1+z^2)^3}$ and C is the contour consisting of the semi circle C_R which is upper half of a large circle $|z| = R$ of radius R together with the part of the real axis from$-R$ to $+R$. For the poles

$$(1+z^2)^3 = 0$$

$$\Rightarrow\ z = \pm i$$

$\therefore z = i$ and $z = -i$ are the two poles each of order 3. But only $z = i$ lies within the C. Residue at $z = i$
To get residue at $z = i$, put $z = i+t$, then

$$\begin{aligned}
\frac{z^2}{(1+z^2)^3} &= \frac{(i+t)^2}{[1+(i+t)^2]^3} = \frac{-1+2it+t^2}{[1-1+2it+t^2]^3} \\
&= \frac{(-1+2it+t^2)}{(2it)^3(1+\frac{1}{2i}t)^3} = \frac{(-1+2it+t^2)}{-8it^3}(1+\frac{t}{2i})^{-3} \\
&= -\frac{1}{8i}\left(-\frac{1}{t^3}+\frac{2i}{t^2}+\frac{1}{t}\right)\left(1-\frac{3t}{2i}+\frac{(-3)(-4)}{2}\,\frac{t^2}{-4}+\ldots\right) \\
&= -\frac{1}{8i}\left[-\frac{1}{t^3}+\frac{2i}{t^2}+\frac{1}{t}\right]\left[1-\frac{3}{2i}t-\frac{3}{2}t^2+\ldots\right]
\end{aligned}$$

Here coefficient of $\frac{1}{t}$ is $\frac{-1}{8i}(\frac{3}{2}-3+1) = \frac{-i}{16}$, which is therefore the residue at $z = i$.

Using Cauchy's theorem of residues we have

$$\int_c f(z)dz = 2\pi i \times Res$$

where Res = Sum of the residues of $f(z)$ at the poles within c.

$$\int_{-R}^{R} f(x)dx + \int_{C_R} f(z)dz = 2\pi i(-\frac{i}{16})$$

$$\int_{-R}^{R} \frac{x^2}{(1+x^2)^3}dx + \int_{C_R} \frac{z^2}{(1+z^2)^3}dz = \frac{\pi}{8}$$

Now

$$\left|\int_{C_R} \frac{z^2}{(1+z^2)^3}dz\right| \le \int_{C_R} \frac{|z|^2\,|dz|}{\left|1+z^2\right|^3} \le \int_{C_R} \frac{|z|^2|dz|}{\left(\left|z^2\right|-1\right)^3}$$

$$\text{Since } z = Re^{i\theta}, |dz| = Rd\theta$$

$$\le \frac{R^2}{(R^2-1)^3}\int_0^{\pi} Rd\theta$$

$$= \frac{\pi R^3}{(R^2-1)^3} \to 0 \text{ as } R \to \infty$$

Hence

$$\int_{-\infty}^{\infty} \frac{x^2}{(1+x^2)^3}dx = \frac{\pi}{8}$$

□

■ EXAMPLE 5.8

Using the complex variables techniques, evaluate the integral

$$\int_0^{\infty} \frac{dx}{x^4+16}$$

Solution: Consider the integral $\int_C f(z)dz$ where $f(z) = \frac{1}{z^4+16}$ and C is the contour consisting of the semi circle C_R which is upper half of a large circle $|z| = R$ of radius R together with the part of the real axis from$-R$ to $+R$.

For the poles

$$z^4+16 = 0 \;\Rightarrow\; z^4 = -16 \;\Rightarrow\Rightarrow\; z^4 = 16e^{i\pi} = 16e^{i(2n+1)\pi} \;\Rightarrow\; z = 2e^{(2n+1)i\pi/4}$$

Now for $n = 0,1,2,3$

$$z_1 = 2e^{i\pi/4}2(cos\frac{\pi}{4} + isin\frac{\pi}{4}) = 2(\frac{1}{\sqrt{2}} + i\frac{1}{\sqrt{2}}) = \sqrt{2}(1+i)$$
$$z_2 = 2e^{3i\pi/4}2(cos\frac{3\pi}{4} + isin\frac{3\pi}{4}) = 2(-\frac{1}{\sqrt{2}} + i\frac{1}{\sqrt{2}}) = \sqrt{2}(-1+i)$$
$$z_3 = 2e^{5i\pi/4}2(cos\frac{5\pi}{4} + isin\frac{5\pi}{4}) = 2(-\frac{1}{\sqrt{2}} - i\frac{1}{\sqrt{2}}) = -\sqrt{2}(1+i)$$
$$z_4 = 2e^{7i\pi/4}2(cos\frac{7\pi}{4} + isin\frac{7\pi}{4}) = 2(\frac{1}{\sqrt{2}} - i\frac{1}{\sqrt{2}}) = \sqrt{2}(1-i)$$

There are four poles , but only two poles at z_1 and z_2 lie within C.

Residue (at $z = 2e^{\frac{i\pi}{4}}$)

$$= \left[\frac{1}{\frac{d}{dz}(z^4+16)}\right]_{z=2e^{\frac{i\pi}{4}}} = \left[\frac{1}{4z^3}\right]_{z=2e^{\frac{i\pi}{4}}} = \frac{1}{4(2e^{i\frac{\pi}{4}})^3} = \frac{1}{32e^{i\frac{3\pi}{4}}} = -\frac{e^{i\frac{\pi}{4}}}{32}$$

Residues (at $z = 2e^{\frac{3i\pi}{4}}$)

$$= \left[\frac{1}{\frac{d}{dz}(z^4+16)}\right]_{z=2e^{\frac{3i\pi}{4}}} = \left[\frac{1}{4z^3}\right]_{z=2e^{\frac{3i\pi}{4}}} = \frac{1}{4(2e^{3i\frac{\pi}{4}})^3} = \frac{1}{32e^{i\frac{9\pi}{4}}} = \frac{e^{-i\frac{\pi}{4}}}{32} \text{ Note Here.}$$

$$\int_c f(z)dz = 2\pi i \times \left(\frac{-e^{i\pi/4} + e^{-i\pi/4}}{32}\right)$$

$$\int_c f(z)dz = -\left(2\pi i \frac{i \sin\frac{\pi}{4}}{16}\right) = \frac{\sqrt{2}\pi}{16}$$

where Res= Sum of residues at poles within C

$$\int_{-R}^{R} f(z)dz + \int_{C_R} f(z)dz == \frac{\sqrt{2}\pi}{16}$$

$$\int_{-R}^{R} \frac{1}{x^4+16}dx + \int_{C_R} \frac{1}{z^4+16}dz = \frac{\sqrt{2}\pi}{16}$$

Now

$$\begin{aligned} \left|\int_{C_R} \frac{1}{z^4+16}dz\right| &\le \int_{C_R} \frac{|dz|}{|z^4+16|} \\ &\le \int_{C_R} \frac{|dz|}{(|z|^4-16)} \\ &\quad \text{Since } z = Re^{i\theta}, |dz| = Rd\theta \\ &\le \int_0^{\pi} \frac{Rd\theta}{R^4-16} = \frac{R\pi}{R^4-16} \to 0 \text{ as } R \to \infty \end{aligned}$$

Hence

$$\int_{-\infty}^{\infty} \frac{1}{x^4+16}dx == \frac{\sqrt{2}\pi}{16}$$

$$2\int_0^{\infty} \frac{1}{x^4+16}dx = \frac{\sqrt{2}\pi}{16}$$

$$\int_0^{\infty} \frac{1}{x^4+16}dx = \frac{\sqrt{2}\pi}{32}$$

□

■ EXAMPLE 5.9

Using complex variables, evaluate the real integral

$$\int_0^{\infty} \frac{cos(3x)dx}{(x^2+1)(x^2+4)}$$

Solution: Consider the integral $\int_C f(z)dz$ where $f(z) = \frac{e^{3iz}}{(z^2+1)(z^2+4)}$ and C is the contour consisting of the semi circle C_R which is upper half of a large circle $|z| = R$ of radius R together with the part of the real axis from$-R$ to $+R$.

For the poles

$$(z^2+1)(z^2+4) = 0 \Rightarrow z^2+1 = 0 \text{ or } z = \pm i \Rightarrow z^2+4 = 0 \text{ or } z = \pm 2i$$

The Poles at $z = i$ and $z = 2i$ lie within the contour.

Residue (at $z = i$)

$$= \lim_{z\to i} \frac{(z-i)e^{3iz}}{(z^2+1)(z^2+4)} = \lim_{z\to i} \frac{e^{3iz}}{(z+i)(z^2+4)} = \frac{e^{-3}}{6i}$$

Residue (at $z = 2i$)

$$= \lim_{z\to 2i} \frac{(z-2i)e^{3iz}}{(z^2+1)(z^2+4)} = \lim_{z\to i} \frac{e^{3iz}}{(z^2+1)(z+2i)} = -\frac{e^{-6}}{12i}$$

By theorem of Residue

$$\int_C f(z)dz = 2\pi i \times Res$$

$$\int_{-R}^{R} \frac{e^{3iz}dz}{(z^2+1)(z^2+4)} + \int_{C_R} \frac{e^{3iz}dz}{(z^2+1)(z^2+4)} = 2\pi i \left[\frac{e^{-3}}{6i} + \frac{e^{-6}}{-12i}\right]$$

$$\lim_{R\to\infty} \int \frac{e^{3izdz}}{(z^2+1)(z^2+4)} = 0, \text{ By Jordan's Lemma}$$

$$\int_{-\infty}^{\infty} \frac{e^{3ix}}{(x^2+1)(x^2+4)} dx = \pi \left[\frac{e^{-3}}{3} - \frac{e^{-6}}{6}\right]$$

$$\int_{-\infty}^{\infty} \frac{cos(3x)dx}{(x^2+1)(x^2+4)} = \text{ Real part of} \int_{-\infty}^{\infty} \frac{e^{3ix}dx}{(x^2+1)(x^2+4)}$$

$$\int_{-\infty}^{\infty} \frac{cos(3x)dx}{(x^2+1)(x^2+4)} = \pi \left[\frac{e^{-3}}{3} - \frac{e^{-6}}{6}\right]$$

$$\int_{0}^{\infty} \frac{cos(3x)dx}{(x^2+1)(x^2+4)} = \frac{\pi}{2} \left[\frac{e^{-3}}{3} - \frac{e^{-6}}{6}\right]$$

□

Problems

5.2 Use Residue of calculus to evaluate the following integrals:

a) $\displaystyle\int_0^\infty \frac{\cos mx}{(x^2+1)} dx$

b) $\displaystyle\int_0^\infty \frac{\log(1+x^2)}{(x^2+1)} dx$

c) $\displaystyle\int_0^\infty \frac{1}{(x^2+1)^3} dx$

d) $\displaystyle\int_0^\infty \frac{1}{(x^6+1)} dx$

e) $\displaystyle\int_0^\infty \frac{\cos x^2 + \sin x^2 - 1}{(x^4+16)} dx$

5.3 By contour integration, prove that $\displaystyle\int_0^\infty \frac{\sin mx}{x} dx = \frac{\pi}{2}$

5.4 Show that, if $a \geq b \geq 0$, then $\displaystyle\int_0^\infty \frac{\cos 2ax - \cos 2bx}{x^2} dx = \pi(b-a)$

5.5 Show that $\displaystyle\int_0^\infty \frac{x^3 \sin x}{(x^2+a^2)(x^2+b^2)} dx = \frac{\pi}{2(a^2-b^2)}[a^2e^{-a} - b^2e^{-b}]$ where $a > b > 0$

UNIT 2
STATISTICAL TECHNIQUES - I

CHAPTER 6

MOMENTS, SKEWNESS AND KURTOSIS

6.1 CONCEPT OF MOMENT

In statistics moment are defined as the mean values of powers of the deviation in any individual series or frequency distribution (discrete and continuous) taken about three points: (i) origin, (ii) mean, or (iii) any other point.

6.1.1 Moments about Origin

6.1.1.1 For an individual series In case of individual series $x_1, x_2, \ldots, x_n$, rth moment about origin is denoted by μ_r' and is defined as

$$\mu_r' = \frac{\sum x^r}{n}$$

where $r = 1, 2, 3, 4,...$

6.1.1.2 For an frequency distribution If $x_1, x_2, ..., x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, ..., f_n$ respectively, then rth moment about origin is denoted by μ_r' and is defined as

$$\mu_r' = \frac{\sum f \times x^r}{\sum f}$$

where $r = 1, 2, 3, 4,...$

6.1.1.3 For an frequency distribution For grouped data, let $x_1, x_2, ..., x_n$ be taken as the mid-values then we have

$$\mu_r' = \frac{\sum f \times x^r}{\sum f}$$

where $r = 1, 2, 3, 4, ...$

6.1.2 Moments about Mean

6.1.2.1 For an individual series In case of individual series $x1, x2, \ldots, xn$, rth moment about mean is denoted by μ_r and is defined as

$$\mu_r = \frac{\sum (x - \overline{x})^r}{n}$$

where $r = 1, 2, 3, 4, ...$

6.1.2.2 For an frequency distribution If $x_1, x_2, ..., x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, ..., f_n$ respectively, then rth moment about mean is denoted by μ_r and is defined as

$$\mu_r = \frac{\sum f \times (x - \overline{x})^r}{\sum f}$$

where $r = 1, 2, 3, 4, ...$

6.1.2.3 For an frequency distribution For grouped data, let $x_1, x_2, ..., x_n$ be taken as the mid- values then we have

$$\mu_r = \frac{\sum f \times (x - \overline{x})^r}{\sum f}$$

where $r = 1, 2, 3, 4, ...$

6.1.3 Moments about Any Arbitrary Point

6.1.3.1 For an individual series In case of individual series $x1, x2, \ldots, xn$, rth moment about any arbitrary point a is denoted by μ_r' and is defined as

$$\mu_r' = \frac{\sum (x - a)^r}{n}$$

where $r = 1, 2, 3, 4, ...$

6.1.3.2 For an frequency distribution If $x_1, x_2, ..., x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, ..., f_n$ respectively, then rth moment about any arbitrary point a is denoted by μ_r' and is defined as

$$\mu_r' = \frac{\sum f \times (x - a)^r}{\sum f}$$

where $r = 1, 2, 3, 4, ...$

6.1.3.3 For an frequency distribution For grouped data, let $x_1, x_2, ..., x_n$ be taken as the mid- values then we have

$$\mu'_r = \frac{\sum f \times (x-a)^r}{\sum f}$$

where $r = 1, 2, 3, 4, ...$

6.2 STANDARD RESULTS

1. Second moment about a is equal to mean square deviation about a., i.e. $\mu'_2(a) = \frac{\sum f \times (x-a)^2}{\sum f}$
2. Second moment about mean is equal to variance, i.e. $\mu_2 = \frac{\sum (x-\overline{x})^2}{\sum f} = \sigma^2$
3. First moment about origin is mean, i.e. $\mu'_1(a) = \frac{\sum fx}{\sum f} = \overline{x}$
4. First moment about mean is always zero, i.e. $\mu_1 = \frac{\sum f \times (x-\overline{x})}{\sum f} = 0$
5. If the series is symmetrical about mean, then all the central moments are zero.
6. The value of a moment of order zero is always 1. i.e., $\mu'_0(a) = \frac{\sum f \times (x-a)^0}{\sum f} = \frac{\sum f}{\sum f} = 1$

6.3 CALCULATION OF MOMENTS ABOUT ORIGIN /ASSUMED MEAN

■ **EXAMPLE 6.1**

The amount in the pockets of 5 students in college were Rs.30,80,110,120 and 200 respectively. Find second moment about origin.

Solution:

x	30	80	110	120	200	Total
x^2	900	6,400	12,100	14,400	40,000	73,800

$$\therefore \mu'_2 = \frac{1}{n}\sum x^2 = \frac{73,800}{5} = 14760 \text{ Rupees}^2$$

□

■ **EXAMPLE 6.2**

Calculate first three moments about assumed mean 89 for the following table:

Marks	81	87	89	90	91	94	96
No. of Students	7	11	15	8	4	3	2

Solution:

x	f	$d = (x - 89)$	fd	fd^2	fd^3
81	7	-8	-56	448	-3584
87	11	-2	-22	44	-88
89	15	0	0	0	0
90	8	1	8	8	8
91	4	2	8	16	32
94	3	5	15	75	375
96	2	7	14	98	686
Total	$N = 50$	-	-33	689	-2571

rth moment about 89

$$\mu_r'(89) = \frac{1}{N}\sum f(x-89)^r, r = 1,2,3.........$$
$$\therefore \mu_1'(89) = \frac{\sum f(x-89)}{\sum f} = \frac{1}{N}\sum fd = \frac{-33}{50} = -6.6$$
$$\mu_2'(89) = \frac{\sum f(x-89)^2}{\sum f} = \frac{1}{N}\sum fd^2 = \frac{689}{50} = 13.78$$
$$\mu_3'(89) = \frac{\sum f(x-89)^3}{\sum f} = \frac{1}{N}\sum fd^3 = \frac{-2571}{50} = 51.42$$

□

6.4 CALCULATION OF CENTRAL MOMENTS

6.4.1 Direct Method to Find Central Moments

It is better to use this method when arithmetic mean is an integer. The first moment about mean $\mu_1 = \dfrac{\sum(x-\overline{x})}{n}$ is always zero.

Steps

1. Calculate mean $\overline{x}$.
2. Obtain deviation $(x-\overline{x})$ or d.
3. Obtain $\sum(x-\overline{x})^2, \sum(x-\overline{x})^3, \sum(x-\overline{x})^4, ...$
4. Obtain rth moment by $\mu_r = \dfrac{\sum(x-\overline{x})^2}{n}$ or $\dfrac{\sum d^r}{n}, r = 1,2,3,4,.......$

In case of discrete and grouped series, the corresponding formula is

$$\mu_r = \frac{\sum f(x-\overline{x})^r}{\sum f} or \frac{\sum fd^r}{\sum f}, r = 1,2,3,4........$$

■ **EXAMPLE 6.3**

Compute the first four central moments for the set 2, 4, 6, 8.:

Solution:

x	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^3$	$(x-\bar{x})^4$
2	-3	9	-27	81
4	-1	1	-1	1
6	1	1	-1	1
8	3	9	27	81
20	0	20	0	164

Mean,

$$\bar{x} = \frac{\sum x}{n} = \frac{20}{4} = 5$$

Central moments

$$\mu_1 = 0, \text{ always}$$
$$\mu_2 = \frac{\sum(x-\bar{x})^2}{n} = \frac{20}{4} = 5$$
$$\mu_3 = \frac{\sum(x-\bar{x})^3}{n} = \frac{0}{4} = 0$$
$$\mu_4 = \frac{\sum(x-\bar{x})^4}{n} = \frac{164}{4} = 41$$

□

■ EXAMPLE 6.4

From the following frequency distribution, compute first four moments about the mean by direct method:

Size(x)	4	8	12	16	20
Frequency(f)	2	2	1	4	1

Solution:

x	f	fx	d	fd	fd^2	fd^3	fd^4
4	2	8	-8	-16	128	-1024	8192
8	2	16	-4	-8	32	-128	512
12	1	12	0	0	0	0	0
16	4	64	+4	+16	64	256	1024
20	1	20	+8	+8	64	512	4096
	$N=10$	$\sum fx = 120$		$\sum fd = 0$	$\sum fd^2 = 288$	$\sum fd^3 = -384$	$\sum fd^4 = 13824$

Mean,

$$\bar{x} = \frac{\sum fx}{N} = \frac{120}{10} = 12$$

The first four moments about mean are:

$$\mu_1 = \frac{0}{10} = 0$$
$$\mu_2 = \frac{\sum fd^2}{N} = \frac{288}{10} = 28.8;$$
$$\mu_3 = \frac{\sum fd^3}{N} = \frac{-384}{10} = -38.4$$
$$\mu_4 = \frac{\sum fd^4}{N} = \frac{13824}{10} = 1382.4$$

□

EXAMPLE 6.5

From the data given below compute the first four moments about the mean by direct method:

Marks	0-10	10-20	20-30	30-40
Frequency	1	2	3	4

Solution:

Marks	Mid-value,x	f	fx	d	fd	fd^2	fd^3	fd^4
0-10	5	1	5	-20	-20	400	-8000	160000
10-20	15	2	30	-10	20	200	-2000	20000
20-30	25	3	75	0	0	0	0	0
30-40	35	4	140	+10	+40	400	+4000	40000
	Σ	10	250		0	1000	-6000	220000

Mean,
$$\bar{x} = \frac{\sum fx}{N} = \frac{250}{10} = 25$$

The first four central moments are:

$$\mu_1 = \frac{\sum fd}{N} = \frac{0}{10} = 0;$$
$$\mu_2 = \frac{\sum fd^2}{N} = \frac{1000}{10} = 100$$
$$\mu_3 = \frac{\sum fd^3}{N} = \frac{-6000}{10} = -600$$
$$\mu_4 = \frac{\sum d^4}{N} = \frac{220000}{10} = 22000$$

□

6.4.2 Indirect Method For Calculating Central Moments

6.4.2.1 Short-cut-Method Steps:

1. Take some convenient values as assumed mean, a,say.

2. Obtain deviations from the assumed mean a i.e., calculate $d_x = x - a$
3. Obtain $fd_x^2, fd_x^3, fd_x^4, \ldots\ldots\ldots\ldots$ and their totals $\sum fd_x, \sum fd_x^2, \sum fd_x^3, \sum fd_x^4, \ldots\ldots\ldots\ldots$
4. Calculate $v_1 = \frac{\sum fd_x}{N}, v_2 = \frac{\sum fd_x^2}{N}, v_3 = \frac{\sum fd_x^3}{N}, v_4 = \frac{\sum fd_x^4}{N}$
5. Finally calculate central moments:

$$\mu_1 = 0, \mu_2 = v_2 - v_1^2, \mu_3 = v_3 - 3v_2v_1 + 2v_1^3, \mu_4 = v_4 - 4v_3v_1 + 6v_2v_1^2 - 3v_1^4$$

■ **EXAMPLE 6.6**

The first four moments about 4 in a distribution are -1.5,17,-30 and 108 respectively.Calculate moments based on mean .

Solution:
Given:

$$\begin{aligned}
v_1 &= \frac{1}{N}\sum f(x-4) = -1.5 \\
v_2 &= \frac{1}{N}\sum f(x-4)^2 = 17 \\
v_3 &= \frac{1}{N}\sum f(x-4)^3 = -30 \\
v_4 &= \frac{1}{N}\sum f(x-4)^4 = 108
\end{aligned}$$

Moments about arithmetic mean:

$$\begin{aligned}
\mu_1 &= 0 \\
\mu_2 &= v_2 - v_1^2 = 17 - (-1.5)^2 = 17 - 2.5 = 14.75 \\
\mu_3 &= v_3 - 3v_2v_1 + 2v_1^3 \\
&= -30 - 3(17)(-1.5) + 2(-1.5)^3 = 39.75 \\
\mu_4 &= v_4 - 4v_3v_1 + 6v_2v_1^2 - 3v_1^4 \\
&= 108 - 4(-30)(-1.5) + 6(17)(-1.5)^2 - 3(-1.5)^4 \\
&= 108 - 180 + 229.5 - 15.1875 = 142.3125
\end{aligned}$$

□

6.4.2.2 Step Deviation Method Steps:

1. Take some convenient value as assumed, say a.
2. Obtain deviations from the assumed mean a i.e, calculate $d_x = x - a$
3. Divide d_x by a suitable value, say h, i.e.,

$$d_s = \frac{d_x}{h} = \frac{x-a}{h}$$

4. Calculate

$$v_1 = \frac{\sum f d_s}{\sum f} = \frac{1}{N} \sum f d_s$$
$$v_2 = \frac{\sum f d_s^2}{\sum f} = \frac{1}{N} \sum f d_s^2$$
$$v_3 = \frac{\sum f d_s^3}{\sum f} = \frac{1}{N} \sum f d_s^3$$
$$v_4 = \frac{\sum f d_s^4}{\sum f} = \frac{1}{N} \sum f d_s^4$$

5. Finally calculate central moments:

$$\mu_1 = 0, \text{ always}$$
$$\mu_2 = (v_2 - v_1^2) \times h^2$$
$$\mu_3 = (v_3 - 3v_2 v_1 + 2v_1^3) \times h^3$$
$$\mu_4 = (v_4 - 4v_3 v_1 + 6v_2 v_1^2 - 3v_1^4) \times h^4$$

■ EXAMPLE 6.7

Calculate first four central moments from the following data:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	1	3	5	7	4

Solution:

Class	Frequency (f)	Mid-value (x)	$d_s = \frac{x-25}{10}$	fd_s	fd_s^2	fd_s^3	fd_s^4
0-10	1	5	-2	-2	4	-8	16
10-20	3	15	-1	-3	3	-3	3
20-30	5	25	0	0	0	0	0
30-40	7	35	1	7	7	7	7
40-50	4	45	2	8	16	32	64
Total	20	-	-	10	30	28	90

Here a = 25,, h = 10 and

$$v_1 = \frac{\sum f d_s}{\sum f} = \frac{10}{20} = 0.5$$
$$v_2 = \frac{\sum f d_s^2}{\sum f} = \frac{30}{20} = 1.5$$
$$v_3 = \frac{\sum f d_s^3}{\sum f} = \frac{28}{20} = 1.4$$
$$v_4 = \frac{\sum f d_s^4}{\sum f} = \frac{90}{20} = 4.5$$

Hence central moments are:

$$\begin{aligned}
\mu_1 &= 0\\
\mu_2 &= (v_2 - v_1^2) \times h^2\\
&= (1.5-(0.5)^2) \times 10^2 = 125\\
\mu_3 &= (v_3 - 3v_2v_1 + 2v_1^3) \times h^3\\
&= \{1.4-3(1.5)(0.5)+2(0.5)^3\} \times 10^3 = -600\\
\mu_4 &= (v_4 - 4v_3v_1 + 6v_2v_1^2 - 3v_1^4) \times h^4\\
&= \{4.5-4(1.4)(.5)+6(1.5)(0.5)^2-3(0.5)^4 \times 10^4 = 37625
\end{aligned}$$

□

6.5 CHARLIER'S CHECK FOR ACCURACY

Charlier has given the following formula which can be applied for testing the accuracy of the following totals while calculating moments.

$$\sum fd_x, \sum fd_x^2, \sum fd_x^3, \sum fd_x^4$$

or

$$\sum fd_s, \sum fd_s^2, \sum fd_s^3, \sum fd_s^4$$

These checks are applied stepwise:

Step	Charlier's Check formula	Up to Moment
First	$\sum f(d_x+1) = \sum fd_x + \sum f$	First
Second	$\sum f(d_x+1)^2 = \sum fd_x^2 + 2\sum fd_x + \sum f$	Second
Third	$\sum f(d_x+1)^3 = \sum fd_x^3 + 3\sum fd_x^2 + \sum fd_x + \sum f$	Third
Fourth	$\sum f(d_x+1)^4 = \sum fd_x^4 + 4\sum fd_x^3 + 6\sum fd_x^2 + 4\sum fd_x + \sum f$	Fourth

6.6 MOMENTS ABOUT ORIGIN IN TERMS OF CENTRAL MOMENTS

If $\mu_1', \mu_2', \mu_3', \mu_4'$ are first four moments about origin then first four moments about mean are:

$$\begin{aligned}
\mu_1' &= \bar{x}\\
\mu_2' &= \mu_2 + (\mu_1')^2 \text{ or } \mu_2 + \bar{x}^2\\
\mu_3' &= \mu_3 + 3\mu_1'\mu_2 + (\mu_1')^3\\
\mu_4' &= \mu_4 + 4\mu_1'\mu_3 + 6(\mu_1')^2\mu_2 + (\mu_1')^4
\end{aligned}$$

■ **EXAMPLE 6.8**

The central moments of distribution whose mean is 5 are 0,3,0 and 26 respectively. Find the moments about origin.

Solution: Given: $\bar{x}=5, \mu_1=0, \mu_2=3, \mu_3=0, \mu_4=26$

$$\begin{aligned}
\therefore \mu_1' &= \bar{x} = 5 \\
\mu_2' &= \mu_2 + (\bar{x})^2 = 3+5^2 = 3+25 = 28 \\
\mu_3' &= \mu_3 + 3\bar{x}\mu_2 + (\bar{x})^3 \\
&= 0+3\times5\times3+5^3 = 45+125 = 170 \\
\mu_4' &= \mu_4 + 4\bar{x}\mu_3 + 6\bar{x}^2\mu_2 \times (\bar{x})^4 \\
&= 26+4\times5\times0+6\times5^2\times3+5^4 = 1101
\end{aligned}$$

□

6.7 MOMENTS ABOUT ORIGIN IN TERMS OF MOMENTS ABOUT ANY POINT

Method I: Convert the moments about a in terms of moments about mean and then obtain moments about origin.

Method II: Obtain by the formula.

■ **EXAMPLE 6.9**

The arithmetic mean of a series is 22 and first four central moments are 0,81,-144 and 14817. Find the first four moments (i) about assumed 25 and (ii) about origin.

Solution:
(i) Moments about assumed mean $a=25$:

$$\begin{aligned}
v_1 &= d = \bar{x}-a = 22-25 = -3 \\
v_2 &= \mu_2 + d^2 = 81+(-3)^2 = 81+9 = 90 \\
v_3 &= \mu_3 + 3\mu_2 d + d^3 \\
&= -144+3\times81\times(-3)+(-3)^3 = -900 \\
v_4 &= \mu_4 + 4\mu_3 d + 6\mu_2 d^2 + 4^4 \\
&= 14817+4\times(-144)\times(-3)+681\times(-3)^2+(-3)^4 = 21000
\end{aligned}$$

(ii) Moments about origin:

$$\begin{aligned}
\mu_1' &= \bar{x} = 22 \\
\mu_2' &= \mu_2 + (\mu_1')^2 = 81+(22)^2 = 81+484 = 565 \\
\mu_3' &= \mu_3 + 3\mu_2(\mu_1) + (\mu_1')^3 \\
&= -144+3\times81\times22+(22)^3 = 15850 \\
\mu_4' &= \mu_4 + 4\mu_3\mu_1' + 6\mu_2(\mu_1')^2 + (\mu_1')^4 \\
&= 14817+4\times(-144)\times(22)+6\times81\times(22)^2+(22)^4 = 471645
\end{aligned}$$

□

■ **EXAMPLE 6.10**

The first four moments of a frequency distribution about the value 4 are -1.5.1.7,-30 and 108. Calculate the moments about origin.

Solution: Let $v'_r = r$th moment about 4. then

$$v'_1 = -1.5, v'_2 = 17, v'_3 = -30, v'_4 = 108$$

Let $\mu'_r = r$th moment about origin, then we have to find: $\mu'_1, \mu'_2, \mu'_3, \mu'_4$. Thus

$$\mu'_1 = \frac{\sum fx}{\sum f} = \frac{\sum f(x-4+4)}{\sum f} = \frac{\sum f(x-4)}{\sum f} + 4\frac{\sum f}{\sum f}$$
$$= v'_1 + 4 = -1.5 + 4 = 2.5$$
$$\mu'_2 = \frac{\sum fx^2}{\sum f} = \frac{\sum f(x-4+4)^2}{\sum f}$$
$$\mu'_2 = \frac{\sum f(x-4)^2}{\sum f} + \frac{2\times 4\sum f(x-4)}{\sum f} + 4^2\frac{\sum f}{\sum f}$$
$$= v'_2 + 8v'_1 + 16 = 17 + 8(-1.5) + 16 = 17 - 12.0 + 16 = 21$$
$$\mu'_3 = \frac{\sum fx^3}{\sum f} = \frac{\sum f(x-4+4)^3}{\sum f}$$
$$= \frac{\sum f(x-4)^3}{\sum f} + \frac{3\times 4\sum f(x-4)^2}{\sum f} + \frac{3\times 4^2\sum f(x-4)}{\sum f} + 4^3\frac{\sum f}{\sum f}$$
$$= v'_3 + 12v'_2 + 48v'_1 + 4^3 = -30 + 12\times 17 + 48(-1.5) + 64 = 166$$
$$\mu'_4 = \frac{\sum fx^4}{\sum f} = \frac{\sum f(x-4+4)^4}{\sum f}$$
$$= \frac{\sum f(x-4)^4}{\sum f} + 4\times 4\frac{\sum f(x-4)^3}{\sum f} + 6\times 4^2\frac{\sum f(x-4)^2}{\sum f} + 4\times 4^3\frac{\sum f(x-4)}{\sum f} + 4^4\frac{\sum f}{\sum f}$$
$$= v'_4 + 16v'_3 + 96v'_2 + 256v'_1 + 256 = 108 + 16(-30) + 96\times 17 + 256(-1.5) + 256 = 1132$$

□

6.8 COEFFICIENT BASED ON CENTRAL MOMENTS

On the basis of relative proportions of different moments, Alpha(α), Beta(β) and Gamma (γ) coefficients are defined by Karl Pearson as follows:

Alpha Coefficient	Beta Coefficient	Gamma Coefficient
$\alpha_1 = \frac{\mu_1}{\sqrt{\mu_2}} = 0$	$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$	$\gamma_1 = \pm\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{\mu_3}{\sigma_3}$
$\alpha_2 = \frac{\mu_2}{\mu_2} = 1$	$\beta_2 = \frac{\mu_4}{\mu_2^2}$	$\gamma_2 = \beta_2 - 3$
$\alpha_3 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$		
$\alpha_4 = \frac{\mu_4}{\mu_2^2}$		

6.9 SHEPPARD'S CORRECTION

Calculation of central moments in a continuous series are based upon the assumption that all the values in a class-interval are concentrated at the midpoint. This assumption leads to a systematic error in the central moments.To correct these errors W.F. Sheppard has given formula:

$$\text{Corrected } \mu_1 = \mu_1 = 0, \text{ no need of correlation}$$

$$\text{Corrected } \mu_2 = \mu_2 \text{ (Calculated)} - \frac{i^2}{12}$$

where i = magnitude of class-interval

$$\text{Corrected } \mu_2 = \mu_3 \text{ (Calculated), no need of correlation}$$

$$\text{Corrected } \mu_4 = \mu_4 \text{ (Calculated)} - \frac{\mu_2(\text{Calculated}) \times i^2}{2} + \frac{7i^4}{240}$$

There is no need of correction in the first and third moments,because in these cases the algebraic signs of the deviations (+,-) remains as they are,hence the error is neutralized because of its compensatory nature . In second and fourth moments, even the negative signs become positive hence error becomes of cumulative nature and need for correction.

Conditions for the application:

1. The class-intervals are equal.
2. The distribution is symmetric or low skewness.
3. The class-intervals is not more than $\frac{1}{20}$ times of the Range.
4. Total frequency is more than 1000.
5. The distribution is not J or V shaped.

■ EXAMPLE 6.11

The second, third and fourth moments are as follows:

$$\mu_2 = 88.75, \mu_3 = -131.25, \mu_4 = 25445.3125$$

If the width of the class -interval is 10, find the correct moments.

Solution:

$$\begin{aligned}
\text{Corrected } \mu_2 &= \mu_2 - \frac{i^2}{12} \\
&= 88.75 - \frac{10^2}{12} = 80.42 \\
\text{Corrected } \mu_3 &= -131.25 (\text{no need for correction}) \\
\text{Corrected } \mu_4 &= \mu_4 - \frac{\mu_2 \times i^2}{2} + \frac{7i^4}{240} \\
&= 25445.3125 - \frac{88.75 \times 100}{2} + \frac{7 \times 10000}{240} \\
&= 25445.3125 - 4437.5 + 291.667 = 21299.4795
\end{aligned}$$

6.10 LIMITATIONS OF MOMENTS

1. Moments of higher order,though important in theory are so extremely sensitive to sampling fluctuations that values calculated for moderate number of observations are quite unreliable and hardly even repay the labor of computation.
2. In case of higher order moments Charlier's checks become difficult.
3. The assumption of normality is rarely found in economic and social studies in practice, so comparison by moments is not appropriate.
4. In case of symmetrical distributions all odd moments ($\mu_1, \mu_3, \mu_5, ...$) about mean are zero.
5. Moments does not exist for some theoretical distributions.

6.11 AIMS/OBJECTS OF MOMENTS

There are four main characteristics of the frequency distributions. The main aim of moments is to measure these characteristics. To measure the four characteristics of frequency distribution first four moments are sufficient:

Moments	Characteristics it measures
First moment about origin	Central tendency
Second moment about origin	Dispersion
Third moment about mean	Skewness
Fourth moment about mean	Kurtosis

To be more specific we consider the uses of moments as follows:

1. *Measure of Central Tendency* : The first moment about origin μ_1' is equivalent to mean. Thus μ_1' is a measure of central tendency.
2. *Measures of Dispersion*: The three measures of dispersion are as follows:

   Mean Deviation: Mean deviation is equivalent to first absolute moment about mean. That is

   $$MD_{\bar{x}} = \frac{1}{\sum f} \sum f[x - \bar{x}] = \text{ First absolute moment about mean}$$

   Standard Deviation: standard deviation is the square root of the mean of the squared deviations of the observations measured from the arithmetic mean . Thus

   $$\sigma = \sqrt{\mu_2}$$

   $S.D.$ = under root of the second moment about mean .

   Variance: The square of $S.D.$ is variance σ^2, which is equal to μ_2.
3. Other measures based on μ_2(or σ^2 or σ) are:

   Probable Error: In terms of standard deviation probable error is given by

   $$PE = \frac{2}{3}\sigma$$

   Modulus: The modulus of individuals series is defined as

   $$\text{Modulus }, C = \sqrt{\frac{2\sum(x - \bar{x})^2}{n}} = \sqrt{2\mu_2}$$

and in case of discrete/continuous series, we have

$$\text{Modulus }, C = \sqrt{\frac{2\sum f(x-\bar{x})^2}{\sum f}} = \sqrt{2\mu_2}$$

Precision: The inverse of the modules is known as Precision. Thus,

$$\text{Precision } = \frac{1}{C} = \frac{1}{\sqrt{2}\mu_2}$$

Fluctuation: Twice of the second central moment or square of the module is known as fluctuation. That is

$$\text{Fluctuation } = (\text{Modulus})^2 = 2\mu_2$$

6.12 SKEWNESS

Skewness is a measure of the degree of asymmetry of a distribution. If the left tail (tail at small end of the distribution) is more pronounced than the right tail (tail at the large end of the distribution), the function is said to have negative skewness. If the reverse is true, it has positive skewness. If the two are equal, it has zero skewness.

6.12.1 Types of Skewness

There are two types of skewness:

1. Positive Skewness
2. Negative Skewness

A distribution is said to be symmetrical if mean, median and mode are identical. If the distribution in which mean is maximum, mode is least and the median lies in between the two, is called a positively skewed distribution. While the distribution in which mode is maximum and mean is least and median lies in between the two, is called negatively skewed distribution.

It is convention that skewness is said to be positive if the large tail of the distribution lies towards the higher values of the variables (as value of variables is always assumed increasing from left to right) and negative in the contrary.

6.12.2 Measure of Skewness

Measures of skewness may be subjective (or abstract or absolute) or relative. Absolute measures tell us the direction and extent of asymmetry in a frequency distribution. Relative measures, often called coefficient of skewness, permit us to compare two or more frequency distributions.

To compare two or more distributions the relative measure of skewness are used. these measures are as follows:

1. Karl Pearson's coefficient of skewness:
2. Bowley's coefficient of skewness:
3. Kelly's coefficient of skewness:
4. Coefficient of skewness based on moments

6.12.3 Karl Pearson's coefficient of skewness

It is denoted by S_{k_p} and is defined as

$$S_{k_p} = \frac{M - M_o}{\sigma}$$

where M denotes mean, M_o denotes mode and σ denotes standard deviation.

In case of ill defined mode for some particular distribution, the value of empirical mode ($M_o = 3M_d - 2M$) is used in the formula. thus

$$S_{k_p} = \frac{M - (3M_d - 2M)}{\sigma} = \frac{3(M - M_{d)}}{\sigma}$$

In both the cases, these coefficients are pure numbers since both numerator and denominator have the same dimensions. The value of the coefficient lies between -3 and +3.

Again, if $M = M_d$, then obviously, skewness is zero and this is interpreted as symmetrical distribution. If $M > M_o$ or $M > M_d$, then the skewness is positive, while if $M < M_o$ or $M < M_d$, then the skewness is negative.

■ EXAMPLE 6.12

From the following data calculate Karl Pearson's coefficient of skewness:

Year	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
Price index of wheat	83	87	93	104	106	109	118	124	126	130

Solution:

S. No	X	$(X - \overline{X})$	$(X - \overline{X})^2$
1	83	-25	625
2	87	-21	441
3	93	-15	225
4	104	-4	16
5	106	-2	4
6	109	1	1
7	118	10	100
8	124	16	256
9	126	18	324
10	130	22	484
Total	$\sum X = 1080$	-	2476

Mean

$$M = \frac{\sum X}{N} = \frac{1080}{10} = 108$$

Median

$$M_d = \frac{\text{5th value + 6th value}}{2} = \frac{106 + 109}{2} = 107.5$$

Standard deviation

$$\sigma = \sqrt{\frac{\sum(X - \overline{X})^2}{N}} = \sqrt{\frac{2476}{10}} = 15.73$$

Hence, Karl Pearson's coefficient is

$$S_{k_p} = \frac{3(M - M_d)}{\sigma} = \frac{3(108 - 107.05)}{15.73} = 0.0954$$

□

■ EXAMPLE 6.13

From the following data find out Karl Pearson's Coefficient of skewness:

Measurement	10	11	12	13	14	15
Frequency	2	4	10	8	5	1

Solution:

Measurement (X)	Frequency (f)	$d_x = X - 12$	dx^2	fd_x	fd_x^2
10	2	-2	4	-4	8
11	4	-1	1	-4	4
12	10	0	0	0	0
13	8	+1	1	+8	8
14	5	+2	4	+10	20
15	1	+3	9	+3	9
	$N = 30$			$\sum fd_x = \pm 13$	$\sum fd_x^2 = 49$

Arithmetic Mean,

$$M = \bar{X} = a + \frac{\sum fd_x}{N} = 12 + \frac{13}{30} = \approx 12.43$$

Mode (by inspection)

$$M_0 = 12$$

Standard Deviation,

$$\sigma = \sqrt{\frac{\sum fd_x^2}{N} - \left(\frac{\sum fd_x}{N}\right)^2}$$

$$= \sqrt{\frac{49}{30} - (\frac{13}{30})^2} \approx 1.2$$

Karl Pearson's Coefficient of Skewness,

$$S_{k_p} = \frac{M - M_o}{\sigma} = \frac{12.43 - 12}{1.2} \approx 0.3583$$

□

■ EXAMPLE 6.14

Calculate Karl Pearson's coefficient of skewness from the following data:

Marks(above)	0	10	20	30	40	50	60	70	80
No.of students	150	140	100	80	80	70	30	14	0

Solution:

Convert the cumulative frequency distribution into ordinary frequency distribution:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80 and above
Students	10	40	20	0	10	40	16	14	0

Now,

Class	ferquency(f)	Mid Value(X)	Cum-frequency	$d_x = X - 25$	d_x^2	fd_x	fd_x^2
0-10	10	5	10	-30	900	-300	9000
10-20	40	15	50	-20	400	-800	16000
20-30	20	25	70	-10	100	-200	2000
30-40	0	35	70	0	0	0	0
40-50	10	45	80	+10	100	+100	1000
50-60	40	55	120	+20	400	+800	16000
60-70	16	65	136	+30	900	+480	14400
70-80	14	75	150	+40	1600	+560	22400
80 and	0	85	150	+50	2500	0	0
	N = 150					$\sum fd_x = 640$	$\sum fd_x^2 = 80800$

Here frequencies are irregular so mode is ill-defined and we use the formula:

$$S_{k_p} = \frac{3(M - M_o)}{\sigma}$$

Arithmetic Mean,

$$M = \bar{X} = a + \frac{\sum fdx}{N} = 35 + \frac{640}{150} \approx 39.27$$

$$\text{Median } M_d = \tfrac{N}{2}th \text{ item} = \tfrac{150}{2}\text{th item} = 75 \text{ item}$$

This is in the class 40-50.

$$M_d = l_1 + \frac{l_2 - l_1}{f}(m - c)$$

$$M_d = 40 + \frac{50 - 40}{10}(75 - 70) = 45$$

Standard Deviation,

$$\sigma = \sqrt{\frac{\sum fd_x^2}{N} - \left(\frac{\sum fd_x}{N}\right)^2}$$

$$\sigma = \sqrt{\frac{80800}{150} - (\frac{640}{150})^2} = \sqrt{538.67 - (4.27)^2} \approx 22.81$$

Coefficient of Skewness,

$$S_{k_p} = \frac{3(M - M_d)}{\sigma} = \frac{3 \times (39.27 - 45)}{22.81} = -0.754$$

□

■ EXAMPLE 6.15

If mode is more than mean by 4.5 and the variance is 121, find coefficient of skewness.

Solution:

$$M - M_o = Mean - Mode = -4.5$$

$$Variance = \sigma^2 = 121$$

$$\therefore \text{Standard deviation } \sigma = \sqrt{121} = 11$$

$$\therefore \text{Coefficient of skewness} = S_{k_p} = \frac{M - M_o}{\sigma}$$

$$S_{k_p} = \frac{-4.5}{11} \approx -0.41$$

□

6.12.4 Bowley's coefficient of skewness

It is defined as

$$S_{k_b} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

where $Q_r$denotes rth quartile and M_d denotes median. The coefficients is pure number and lies between -1 and +1.

The coefficient is positive or negative if $Q_3 + Q_1 > 2M_d$ or $Q_3 + Q_1 < 2M_d$ respectively. Again, it is zero if $Q_3 + Q_1 = 2M_d$, which shows the distribution is symmetric.

EXAMPLE 6.16

Compute an appropriate measures of skewness for the following data:

Sales (Rs Lakhs)	Below 50	50-60	60-70	70-80	80-90	90-100	100-110	110-120	Above 120
No. of Companies	12	30	65	78	80	55	45	25	10

Solution: For open ended series, most suitable skewness measure is Bowley's Coefficient of Skewness. To compute this we need quartiles.

Sales (Rs Lakhs)	No of Companies	Cumulative Frequencies
Below 50	12	12
50-60	30	42
60-70	65	107
70-80	78	185
80-90	80	265
90-100	55	320
100-110	45	365
110-120	25	390
Above 120	10	400
Total	$N = \sum f = 400$	

rth quartile is given as

$$Q_r = l + \frac{\left[r \times \frac{N}{4} - C\right]}{f} \times h$$

where N denotes total of frequencies, Now we choose quartile class corresponding to the value $r \times \frac{N}{4}$. Then l denotes lower limit of quartile class, f frequency of quartile class, C cumulative frequency just before the quartile class and h denotes interval.

Now, for Q_1, since $\frac{N}{4} = 100$, the quartile class is 60-70.

$$\therefore l = 60,\ C = 42,\ h = 10,\ f = 65$$

$$\begin{aligned} Q_1 &= l + \frac{\left[\frac{N}{4} - C\right]}{f} \times h \\ &= 60 + \frac{100 - 42}{65} \times 10 \\ &= 67.38 \end{aligned}$$

For Q_2, since $2 \times \frac{N}{4} = 200$, the quartile class is 80-90.

$$\begin{aligned} Q_2 &= l + \frac{\left[2 \times \frac{N}{4} - C\right]}{f} \times h \\ &= 80 + \frac{200 - 185}{80} \times 10 \\ &= 81.88 \end{aligned}$$

For Q_3, since $3 \times \frac{N}{4} = 300$, the quartile class is 90-100.

$$\begin{aligned} Q_3 &= l + \frac{\left[3 \times \frac{N}{4} - C\right]}{f} \times h \\ &= 90 + \frac{300 - 265}{55} \times 10 \\ &= 96.36 \end{aligned}$$

Hence, Bowley's Coefficient of Skewness

$$S_{k_b} = \frac{Q_3 + Q_1 - 2Q_1}{Q_3 - Q_1} = \frac{96.36 + 67.38 - 2 \times 81.88}{96.36 - 67.38} = -0.00069$$

This shows that there is a negative skewness, which has a very negligible magnitude. □

■ EXAMPLE 6.17

Calculate coefficient of skewness using quartiles:

Mid-value	15	20	25	30	35	40
Frequency	30	28	25	24	10	21

Solution:

Here, we will convert mid-values in class interval form.

Class (Exclusive form)	Frequency f	Cumulative Frequency
12.5-17.5	30	30
17.5-22.5	28	58
22.5-27.5	25	83
27.5-32.5	24	107
32.5-37.5	10	117
37.5-42.5	21	138
Total	138	

$$q_1 = \frac{N}{4} = \frac{138}{4} = 34.5$$

$$q_2 = \frac{2N}{4} = \frac{2 \times 138}{4} = 69.0$$

$$q_3 = \frac{3N}{4} = \frac{3 \times 138}{4} = 103.5$$

First Quartile,

$$Q_1 = l_1 + \frac{q_1 - c}{f} \times i$$

$$= 17.5 + \frac{34.5 - 30}{28} \times 5 = 18.3$$

Second Quartile,

$$Q_2 = l_1 + \frac{q_2 - c}{f} \times i$$

$$= 22.5 + \frac{69 - 58}{25} \times 5 = 24.7$$

Third Quartile,

$$Q_3 = l_1 + \frac{q_3 - c}{f} \times i$$

$$= 27.5 + \frac{103.5 - 83}{24} \times 5 = 31.8$$

Coefficient of skewness,

$$S_{k_b} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

$$= \frac{31.8 + 18.3 - 2 \times 24.7}{31.8 - 18.3} = \approx 0.052$$

□

■ EXAMPLE 6.18

If Bowley coefficient of skewness is-0.36,$Q_1 = 8.6$ and $M = 12.3$, then find Q_3 and Quartile Coefficient of Dispersion .

Solution:

Bowley coefficient of skewness,

$$S_{k_b} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

$$\therefore -0.36 = \frac{Q_3 + 8.6 - 2 \times 12.3}{Q_3 - 8.6}$$

$$\Rightarrow -0.36(Q_3 - 8.6) = Q_3 + 8.6 - 24.6$$

$$\Rightarrow -1.36Q_3 = -19.096$$

$$\Rightarrow Q_3 = \frac{19.096}{1.36} = 14.04$$

Quartile Coefficient of Dispersion

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{14.04 - 8.6}{14.04 + 8.6} = \frac{5.44}{22.64} = 0.24$$

□

6.12.5 Kelly's coefficient of skewness

It is defined as

$$S_k = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}$$

and

$$S_k = \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

where $D_r$denotes rth decile and P_r denotes rth percentile.

■ EXAMPLE 6.19

From the following data, calculate the measure of skewness using the percentiles

Marks (less than)	10	20	30	40	50	60
No. of students	4	10	30	40	47	50

Solution:

Class interval	Frequency(f)	Cumulative frequency (CF)
0-10	4	4
10-20	6	1
20-30	20	30
30-40	10	40
40-50	7	47
50-60	3	50

$$D_1 = l + \frac{[(N/10) - C]}{f} \times h$$

Here $N = 50$, since CF is just greater than $\frac{N}{10}$ is 10 therefore,

$$l = 10, C = 4, h = 10, \ f = 6$$

$$D_1 = 10 + \frac{(5-4)}{6} \times 10 \approx 11.67$$

$$D_9 = l + \frac{[(9N/10) - C]}{f} \times h$$

Since CF is just greater than $9 \times \frac{N}{10} = 45$ is 47, therefore

$$l = 40, C = 40, h = 10, \ f = 7$$

$$D_9 = 40 + \frac{(45-40)}{7} \times 10 \approx 47.14$$

$$D_5 = l + \frac{[(5N/10) - C]}{f} \times h$$

Since CF is just greater than $5 \times \frac{N}{2} = 25$ is 30, therefore

$$l = 20, C = 10, h = 10, \ f = 20$$

$$D_5 = 20 + \frac{(25-10)}{20} \times 10 = 27.50$$

Kelly's coefficient of skewness

$$S_k = \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

$$= \frac{47.14 + 11.67 - (2 \times 27.50)}{41.14 - 11.67} = \frac{3.81}{35.47} \approx 0.1074$$

□

6.12.6 Measure of Skewness Based on Moments

We measure skewness with the help of moments as follows:
if $\mu_3 = 0$, no skewness
if $\mu_3 > 0$, Positive skewness
if $\mu_3 < 0$,negative skewness
with this absolute measure of skewness we have two coefficients of skewness; Beta Coefficient (β_1) and Gamma Coefficient (γ_1):

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\gamma_1 = \pm\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{\mu_3}{\sigma_3}$$

the sign of $\gamma_1$depends on μ_3.
If $\gamma_1 = 0$, no skewness
$\gamma_1 > 0$, positive skewness
$\gamma_1 < 0$,negative skewness

Remark:

- β_1 does not give the nature of skewness (whether positive or negative) but we consider the sign of μ_3.

- Another coefficient of skewness based on moments is

$$\frac{\sqrt{\beta_1}(\beta_2+3)}{2(5\beta_2-6\beta_1-9)}$$

■ EXAMPLE 6.20

Calculate coefficient of skewness based on moments for the following frequency distribution:

Years	Frequency
2.5- 7.5	8
7.5-12.5	15
12.5-17.5	20
17.5-22.5	32
22.5-27.5	23
27.5-32.5	17
32.5-37.5	5

Solution:

Class	Frequency f	Mid-value x	$d_s=\frac{x-20}{5}$	fd_s	fd_s^2	fd_s^3
2.5-7.5	8	5	-3	-24	72	-216
7.5-12.5	15	10	-2	-30	60	-120
12.5-17.5	20	15	-1	-20	20	-20
17.5-22.5	32	20	0	0	0	0
22.5-27.5	23	25	1	23	23	23
27.5-32.5	17	30	2	34	68	136
32.5-37.5	5	35	3	15	45	135
Total	$\sum f = 120$			$\sum fd_s = -2$	$\sum fd_s^2 = 288$	$\sum fd_s^3 = -62$

$$\mu_1' = \frac{\sum fd_s}{\sum f} = \frac{-2}{120} = -0.0167 = -0.02$$

$$\mu_2' = \frac{\sum fd_s^2}{\sum f} = \frac{288}{120} = 2.4$$

$$\mu_3' = \frac{\sum fd_s^3}{\sum f} = \frac{-62}{120} = -0.5167 = -0.52$$

$$\text{Now } \because \mu_2 = \{\mu_2' - (\mu_1')^2\}i^2, d_s = \frac{x-a}{i}$$

$$= \{2.4-(-0.2)^2\}\times 5^2 = 2.3996\times 5^2$$

$$\therefore \mu_2^3 = 13.824\times 5^6$$

$$\therefore \mu_2^{\frac{3}{2}} = \sqrt{13.824}\times 5^3 = 3.718\times 5^3$$

and

$$\begin{aligned}\mu_3 &= \{\mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3\} \times i^3 \\ &= \{-.52 - 3 \times 2.4 \times (-.02) + 2 \times (-.02)^3\} \times 5^3 = -.376 \times 5^3 \\ \gamma_1 &= \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{-.376 \times 5^3}{3.718 \times 5^3} = \frac{-.376}{3.718} = -0.101\end{aligned}$$

In this frequency distribution the skewness is negative. □

6.13 KURTOSIS

Definition 7. *Kurtosis refers to the extent to which unimodal frequency curve is peaked.*

In other words,

Definition 8. *Kurtosis is a measure that refers to the peakedness of the top of the curve . Kurtosis gives the degree of flatness or peakedness in the region about the mode of a frequency distribution.*

According to Croxton and Cowden,

Definition 9. *A measure of kurtosis indicates the degree to which the curve of a frequency distribution is peaked or flat topped.*

According to Clark and Sckkade,

Definition 10. *Kurtosis is the property of a frequency distribution which expresses its relative peakedness.*

Karl Pearson in 1905 introduced the three types of curves on the basis of kurtosis:

Mesokurtic: If the concentration of frequency in the middle of the frequency distribution is normal, the curve is known as mesokurtic.

Leptokurtic: If the frequencies are densely concentrated in the middle of the series, the will be more peaked than normal and is known as Leptokurtic.

Platykurtic: If the frequencies are not densely concentrated in the middle of the series, the curve will be more flat than normal and is known as platykurtic.

6.14 MEASURE OF KURTOSIS

The measure of kurtosis based on central moments are given by Karl Pearson:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

if $\beta_2 = 3$, the curve is Mesokurtic or normal
if $\beta_2 > 3$, the curve is Leptokurtic or more peaked
if $\beta_2 < 3$, the curve is Platykurtic or flat topped.
The measure of Kurtosis is also represented by gamma two, $\gamma_2 = \beta_2 - 3$
if $\gamma_2 = 0$,the curve is Mesokurtic
if $\gamma_2 > 0$,the curve in Leptokurtic
if $\gamma_2 < 0$, the curve is Platykurtic.

EXAMPLE 6.21

the first four central moments of a distribution are 0,2.5,0.7,and 18.75. Examine the skewness and kurtosis .

Solution:

Given: First four central moments

$$\mu_1 = 0, \mu_2 = 2.5, \mu_3 = 0.7, \mu_4 = 18.75$$

Coefficient of skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = \frac{0.49}{15.625} = 0.3136$$

$$\gamma_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{0.7}{(2.5)^{\frac{3}{2}}} = \frac{0.7}{\sqrt{15.625}} = \frac{0.7}{3.95} = 0.177$$

The distribution is positively skew.
Coefficient of Kurtosis:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = \frac{18.75}{6.25} = 3$$

The curve is mesokurtic.

□

EXAMPLE 6.22

Calculate β_2 from the following data and give your comments on the shape of the distribution:

$$N = 100, \sum f d_x = 75, \sum f d_x^2 = 4000, \sum f d_x^4 = 76500$$

Solution:

$$\mu_1' = \frac{\sum f d_x}{N} = \frac{75}{100} = 0.75$$

$$\mu_2' = \frac{\sum f d_x^2}{N} = \frac{1800}{100} = 18$$

$$\mu_3' = \frac{\sum f d_x^3}{N} = \frac{4000}{100} = 40$$

$$\mu_4' = \frac{\sum f d_x^4}{N} = \frac{76500}{100} = 765$$

Central Moments:

$$\mu_1 = 0$$
$$\mu_2 = \mu_2' - (\mu_1')^2 = 18 - (0.75)^2 = 18 - .5625 = 17.4375$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$
$$= 40 - 3 \times 18 \times 0.75 + 2 \times (0.75)^3 = 0.3438$$
$$\mu_4 = \mu_4' - 4\mu_2'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$
$$= 765 - 4 \times 40 \times .75 + 6 \times 18 \times (.75)^2 - 3(.75)^4 = 704.8008$$
$$\therefore \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(.3438)^2}{(17.4375)3} = \frac{.1182}{5302.158} = .002$$

The distribution may be considered symmetrical.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{704.8008}{(17.4375)^2} = 2.32$$

The distribution is platykurtic. □

■ EXAMPLE 6.23

Calculate first four central moments and coefficient of kurtosis for the following distribution and comment on the result :

Class	3-6	6-9	9-12	12-15	15-18	18-21	21-24
Frequency	28	292	389	212	59	18	2

Solution:

Class	Frequency f	Mid-value x	$d_s = \frac{x-10.5}{3}$	fd_s	fd_s^2	fd_s^3	fd_s^4
3-6	28	4.5	-2	-56	112	-224	448
6-9	292	7.5	-1	-292	292	-292	292
9-12	389	10.5	0	0	0	0	0
12-15	212	13.5	1	212	212	212	212
15-18	59	16.5	2	118	236	472	944
18-21	18	19.5	3	54	162	486	1458
21-24	2	22.5	4	8	32	128	512
Total	1000			44	1046	782	3866

$$\mu_1' = .044, \mu_2' = 1.046, \mu_3' = .782, \mu_4' = 3.866$$

$$\mu_2 = [\mu_2' - (\mu_1')^2] \times i^2$$

$$= [1.046 - (.044)^2] \times 9 = 9.396$$

$$\mu_3 = [\mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 \times i^3 = 17.39$$

$$\mu_4 = [\mu_1' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4] \times i^4$$

$$= [3.866 - 4(.782)(.044) + 6(1.046)(.044)^2 - 3(.044)^4] \times 81$$

$$= 3.74 \times 81 = 302.94$$

$$\therefore \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{302.94}{(9.396)^2} = 3.43, \beta_2 > 3,$$

and

$$\gamma_2 = \beta_2 - 3 = 3.43 - 3 = 0.43 > 0,$$

□

Problems

6.1 Compute the first four moments about mean for: x: 14, 16, 18, 20, 25, 27.

6.2 Compute the first three moments about 25:

Expenses	0-10	10-20	20-30	30-40	40-50
No. of Persons	1	2	5	2	1

6.3 Find the first four moments about the mean from the following data:

Salary	0-5	5-10	10-15	15-20	20-25
No. of Employees	1	4	6	8	9

6.4 Calculate skewness and kurtosis for the following data:

Size	0	10	20	30	40
Frequency	10	8	6	3	1

6.5 Find the kurtosis coefficient for the following frequency distribution:

Marks	5-15	15-25	25-35	35-45	45-55
No. of students	1	3	5	7	4

6.6 Calculate the first four moments of the following distribution about the mean and hence find $\beta_1$and β_2:

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

6.7 The first three moments about the value 3 are 2,10 and 30 respectively. Obtain the first three momens about zero.

6.8 The first four moments of a distribution about x = 4 are 1,4,10 and 45 respectively. Show that mean is 5. Also calculate the central moments β_1 and β_2.

6.9 The first four moments of a distribution about the value 2 are 1,2.5,5.5, and 1.6. Calculate the moments about the mean and origin.

6.10 The first four moments about assumed mean of a frequency distributions are $v_1 = -2, v_2 = 8, v_3 = -20, v_4 = 60$. Find the values of β_1 and β_2.

6.11 The first four moments of a distribution about the assumed mean 5 are 2,20,40, and 150 respectively .Find central moments, β_1 and β_2.

6.12 Calculate Bowley's Coefficient of skewness and γ_1 from the following data:

Income in Rs.	No. of persons
Less than 500	15
Less than 600	28
Less than 700	45
Less than 800	70
Less than 900	100
Less than 1000	120
Less than 1100	130

6.13 Calculate Karl Pearson's coefficient of skewness from the following data:

Class	0-6	6-12	12-18	18-24	24-30	30-36
Frequency	12	24	38	52	34	19

6.14 Calculate Karl Pearson's coefficient of skewness from the data given below:

Life Time (hour)	No. of Tubes
300-400	14
400-500	46
500-600	58
600-700	76
700-800	68
800-900	62
900-1000	48
1000-1100	22
1100-1200	6

6.15 Calculate mean, standard deviation and coefficient of skewness from the following data:

Class	40-45	45-50	50-55	55-60	60-65	65-70	70-75	75-80
Frequency	4	12	20	35	32	16	3	1

6.16 Calculate Karl Pearson's coefficient of skewness from the following data:

Marks	No. of students
less than 20	8
less than 30	38
less than 40	53
40-50	2
50-60	8
60-70	30
more than 70	17
more than 80	4

6.17 Find Bowley's coefficient of skewness from the following data:

Measurement	10	11	12	13	14	15
Frequency	2	4	10	8	5	1

6.18 calculate coefficient of skewness based on quartiles:

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Frequency	4	10	12	4	6	16	8

6.19 Find out quartile coefficient of skewness from the following data:

Marks	5-15	15-25	25-35	35-45	45-55	55-65
Frequency	5	15	35	55	25	15

6.20 Calculate quartiles from the following data and find coefficient of skewness based on quartiles:

Class-interval	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
Frequency	5	9	14	20	25	15	8	4

6.21 Find the coefficient of skewness of the following two groups A and B and point out which distribution is more skewed ?

Marks	55-58	58-61	61-64	64-67	67-70
Group A	12	17	23	18	11
Group B	20	22	25	13	7

6.15 RANDOM VARIABLE

In an experiment of tossing a pair of coin we may be intrested in obtaining the probabilities of getting 0,1 or 2 heads which are possible events. We consider a variable X which hold these possible values 0,1 or 2. This variable is a real valued function defined over sample space whose range is non empty set of real numbers is called random variable. i.e.,

A random variable is a function $X(w)$ with domain S and range $(-\infty,\infty)$ such that for every real number a, the event $[e : X(e) \leq a]$.

6.16 PROBABILITY MASS FUNCTION (PMF)

If X is a discrete random variable with distinct values $x_1, x_2, \ldots, x_n$ for which X has positive probabilities $p_1, p_2, \ldots, p_n$, then the function $f(x)$defined as:

$$f_X(x) = \begin{cases} p_i & \text{if } x = x_{i,} \\ 0 & \text{if } x \neq x_i \end{cases}$$

is called the probability mass function of random variable X.

Since $P(S) = 1$(Probability of sure event is one.), we must have

$$\sum_j f(x_j) = 1$$

6.17 PROBABILITY DENSITY FUNCTION (PDF)

Similar to PMF, probability density function is defined for continuous random variable X. The funxtion $f(x)$ defined as

$$P(x \leq X \leq x + dx) = f(x)dx$$

such that

- $f(x) \geq 0$ for any x
- $\int_{-\infty}^{\infty} f(x)dx = 1$

6.18 MATHEMATICAL EXPECTATION OR EXPECTED VALUE OF A RANDOM VARIABLE.

The expected value of a discrete random variable is a weighted average of all possible values of the random variable, where the weights are the probabilities associated with the corresponding values.

For discret random variable. The expected value of a discrete random variable X with probability mass function (p.m.f)$f(x)$is given below:

$$E(X) = \sum_x x f(x)$$

For continuous random variable. The mathematical expression for computing the expected value of a continuous random varriable X with probability density function (p.d.f.) f(x) is, however, as follows:

$$E(X) = \int_{-\infty}^{\infty} x f(x)dx$$

Particular Case:

$$E(X^r) = \int_{-\infty}^{\infty} x^r f(x)dx,$$

which is defined as μ_r' the rth moment (about origin) of the probability distribution. Thus

$$\mu_r'\text{(about origin)} = E(X^r).$$

$$\mu_1'\text{(about origin)} = E(X)$$

$$\mu_2'(\text{about origin}) = E(X^2)$$

Hence, Mean$=\bar{x}=\mu_1'$(about origin)$=E(X)$ and $\mu_2=\mu_2'-\mu_1'^2=E(X^2)-\{E(X)\}^2$

$$\mu_2 = E[X-E(X)]^2 = \int_{-\infty}^{\infty}(x-\bar{x})^2 f(x)dx$$

Remark:

- $E(c)=\int_{-\infty}^{\infty} cf(x)dx = c\int_{-\infty}^{\infty} f(x)dx = c$

6.18.1 Properties of Expectation

1. Addition theorem: If X and Y are random variables, then $E(aX+bY)=aE(X)+bE(Y)$ and we can generalise above result as $E(\sum_{i=1}^{n} X_i)=\sum_{i=1}^{n} E(X_i)$, if all the expectations exist.
2. Multiplication theorem: If X and$Y$ are random variables, then $E(XY)=E(X)E(Y)$ provided X and Y are independent and we can generalise above result as $E(\prod_{i=1}^{n} X_i)=\prod_{i=1}^{n} E(X_i)$, if all the expectations exist.
3. Expectation of a linear Combination of Random Variable: Let $X_1,X_2,\ldots,X_n$ be any n random variables and if $a_1,a_2,\ldots,a_n$ are any n constants, then $E(\sum_{i=1}^{n} a_iX_i)=\sum_{i=1}^{n} a_iE(X_i)$
4. If $X\geq 0$ then $E(X)\geq 0$.
5. If X,Y are two random variable s.t.$Y\leq X$, then $E(Y)\leq E(X)$
6. $|E(X)|\leq E(|X|)$

6.19 MOMENT GENERATING FUNCTION

The function $E(e^{tX})$ serves to generate moments of the probability distribution of the variate about the origin. and is called Moment Generating Function.

The Moment Generating Function (m.g.f)of a random variable X(about origin) having the probability function $f(x)$

$$M_X(t)=E(e^{tX})=\begin{cases}\int e^{tx}f(x)dx & \text{For Continuous}\\ \sum_x e^{tx}f(x) & \text{For Discrete}\end{cases}$$

We consider for discrete distribution,

$$\begin{aligned}M_X(t)=E(e^{tx})&=E(1+tX+\frac{t^2X^2}{2!}+\ldots\ldots\ldots.+\frac{t^rX^r}{r!}+\ldots\ldots\ldots\ldots.)\\ &=1+tE(X)+\frac{t^2E(X^2)}{2!}+\ldots\ldots\ldots.+\frac{t^rE(X^r)}{r!}+\ldots\ldots\ldots\ldots.\\ &=1+t\mu_1'+\frac{t^2\mu_2'}{2!}+\ldots\ldots\ldots.+\frac{t^r\mu_r'}{r!}+\ldots\ldots\ldots\ldots.\\ &=\sum_{r=0}^{\infty}\frac{t^r}{r!}\mu_r'\end{aligned}$$

Similarly, we can show for continuous distribution. Thus rth moment of Xabout origin

$$\mu_r'=E(X^r)=\begin{cases}\int x^r f(x)dx & \text{For Continuous}\\ \sum_x x^r p(x)dx & \text{For Discrete}\end{cases}$$

We see that the coefficient of $\frac{t^r}{r!}$in $M_X(t)$ gives μ_r' (about origin).

On differentiating above equation with respect to r and then putting$t = 0$, we get

$$\frac{d^r}{dt^r}M_X(t)\Big|_{t=0} = \left[\frac{\mu_r'}{r!}.r! + \mu_{r+1}^{'}t + \mu_{r+2}^{'}\frac{t^2}{2!} + \dots\right]_{t=0}$$

$$\mu_r^{'} = |\frac{d^r}{dt^r}M_X(t)|_{t=0}$$

In general, the moment generating function X about the point $X = a$ is defined as

$$\begin{aligned} M_{X-a}(t)(\text{About } X = a) &= E[e^{t(x-a)}] \\ &= E\{1 + t(X-a) + \frac{t^2}{2!}(X-a)^2 + \dots + \frac{t^r}{r!}(X-a)^r + \dots \\ &= 1 + t\mu_1^{'} + \frac{t^2}{2!}\mu_2^{'} + \dots + \frac{t^r}{r!}\mu_r^{'} + \dots \end{aligned}$$

where $\mu_r' = E\{(X-a)^r\}$, is the$r$th moment about the point $X = a$

6.19.1 Properties of Moment Generating Function(m.g.f)

1. $M_{cX}(t) = M_X(ct)$ c being constant.
2. If $X_1, X_2, X_3, \dots\dots\dots\dots\dots\dots X_n$are independent random variables, then the moment generating function of their sum $X_1 + X_2 + \dots\dots + X_n$ is given by:

$$M_{X_1+X_2+\dots\dots+X_n}(t) = X_1(t)M_{X_2}(t)\dots\dots\dots\dots M_{X_n}(t)$$

3. Effect of change of origin and scale on m.g.f: Let ustransform X to the new variable U by changing both the origin and scale in X as follows: $U = \frac{X-a}{h}$, where a and h are constants. M.G.F.of U (about origin) is given by :

$$M_U(t) = E\left(e^{tU}\right) = E[e^{\left(\frac{t(x-a)}{h}\right)}] = E[e^{t\frac{X}{h}}e^{-\frac{at}{h}}] = e^{-\frac{at}{h}}E\left(e^{\frac{tX}{h}}\right) = e^{-\frac{at}{h}}E\left(e^{\frac{Xt}{h}}\right) = e^{\frac{-at}{h}}M_X\left(\frac{t}{h}\right)$$

where $M_X(t)$ is the m.g.f of X about origin.

In particular $a = E(X) = \mu, h = \sigma_x = \sigma$ then

$$U = \frac{X - E(X)}{\sigma_X} = \frac{X-\mu}{\sigma} = Z$$

is known as a standard variate. Thus the m.g.f of a standard variate Z is given by

$$M_Z(t) = e^{\frac{-\mu t}{\sigma}}M_X(\frac{t}{\sigma})$$

Remark

- $E(Z) = E(\frac{X-\mu}{\sigma}) = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}\{E(X) - \mu\} = \frac{1}{\sigma}(\mu - \mu) = 0$
- $V(Z) = 1$
- If moment generating function exist it is unique.

■ **EXAMPLE 6.24**

Let the random variable X assume the value r with the probability law:

$$P(X = r) = q^{r-1}p; r = 1, 2, \dots$$

Find the m.g.f of X and hence its mean and variance.

Solution:

$$
\begin{aligned}
M_X(t) = E(e^{tx}) &= \sum_{r=1}^{\infty} e^{tr} q^{r-1} p \\
&= \frac{p}{q} \sum_{r=1}^{\infty} \left(qe^t\right)^r = pe^t \sum_{r=1}^{\infty} \left(qe^t\right)^{r-1} \\
&= pe^t \left[1 + qe^t + (qe^t)^2 + \dots\right] = \left(\frac{pe^t}{1-qe^t}\right)
\end{aligned}
$$

$$
M'_X(t) = \frac{pe^t}{(1-qe^t)^2}
$$

$$
M''_X(t) = pe^t \frac{(1+qe^t)}{(1-qe^t)^3}
$$

μ'_1(about origin)$= M'_X(0) = \frac{p}{(1-q)^2} = \frac{1}{p}$
μ'_2(about origin)$=M''_X(0) = \frac{p(1+q)}{(1-q)^3} = \frac{1+q}{p^2}$
Hence, mean $= \mu'_1 = \frac{1}{p}$ and variance $= \mu_2 = \mu'_2 - \mu'^2_1 = \frac{q}{p^2}$

□

■ EXAMPLE 6.25

If the moments of variate X are defined by $E(X^r) = 0.6$; $r = 1,2,3,\dots$. Show that $P(X = 0) = 0.4$, $P(X = 1) = 0.6$ $P(X \geq 2) = 0$

Solution:
The m.g.f. of variate X is:
M_X

$$
M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu_{r'} = 0.4 + 0.6e^t
$$

$$
\begin{aligned}
M_X(t) = E(e^{tX}) &= \sum e^{tx} P(X=x) \\
&= P(X=0) + e^t P(X=1) + \sum_{x=2}^{\infty} e^{tx} P(X=x)
\end{aligned}
$$

on comparing we get the answer.

□

■ EXAMPLE 6.26

Find the m.g.f. of the random variable whose moments are :

$$
\mu_r^{'} = (r+1)! 2^r
$$

Solution:

$$\begin{aligned} M_X(t) &= \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu_r' = \sum_{r=0}^{\infty} \frac{t^r}{r!}(r+1)!2^r \\ &= (1-2t)^2 \end{aligned}$$

□

Problems

6.22 Define m.g.f of a r.v. Find m.g.f of (i)$Y = aX + b$, (ii)$Y = \frac{X-m}{\sigma}$

6.23 Show that if $\overline{X}$ is mean of n random variables:$M_{\overline{X}}(t) = \left[M_X\left(\frac{t}{n}\right)\right]^n$

CHAPTER 7

CURVE FITTING

The curve fitting process fits the equation of approximating curves to the raw field data. Nevertheless, for a given set of data, the fitting curves of a given type are generally NOT unique. Thus, a curve with a minimal deviation from all data points is desired.

7.1 THE METHOD OF LEAST SQUARES

Principle of Least Square: The curve of best fit is that for which the sum of squares of the residuals is minimum.

The method of least squares assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (least square error) from a given set of data.

Let the set of data points be $(x_i, y_i), i = 1, 2, \ldots, m$, and let the curve given by $y = f(x)$ be fitted to this data. At $x = x_i$, the observed value of the ordinate is y_i and the corresponding value on the fitting curve is $f(x_i)$. If ε_i is the error of approximation at $x = x_i$, then we have

$$\varepsilon_i = y_i - f(x_i) \tag{7.1}$$

The sum of square of errors,

$$\varepsilon = \Sigma\varepsilon_i^2 = \Sigma[y_i - f(x_i)]^2 \tag{7.2}$$

Now if $\varepsilon = 0$ then $y_i = f(x_i)$, i.e. all the points lie on the curve. Otherwise, the minimum ε results the best fitting of the curve to the data.

7.1.1 Fitting of A Straight Line

Let (x_i, y_i) be n sets of observations of related data and $y = a + bx$ be the straight line to be fitted. Then from equation 7.2,

$$\varepsilon = \Sigma[y_i - f(x_i)]^2 = \Sigma[y_i - (a + bx_i)]^2$$

By the principle of least square, ε is minimum for some a and b, i.e.,

$$\frac{\partial \varepsilon}{\partial a} = 0 \qquad \text{and} \qquad \frac{\partial \varepsilon}{\partial b} = 0$$

$$\begin{aligned} \frac{\partial \varepsilon}{\partial a} = 0 \qquad \Rightarrow \qquad & \Sigma 2[y_i - (a + bx_i)](-1) = 0 \\ & \Rightarrow \Sigma[y_i - (a + bx_i)] = 0 \\ & \Rightarrow \Sigma y_i = na + b\Sigma x_i \end{aligned} \tag{7.3}$$

Similarly,

$$\begin{aligned} \frac{\partial \varepsilon}{\partial b} = 0 \qquad \Rightarrow \qquad & \Sigma 2[y_i - (a + bx_i)](-x_i) = 0 \\ & \Rightarrow \Sigma[y_i x_i - (ax_i + bx_i^2)] = 0 \\ & \Rightarrow \Sigma y_i x_i = a\Sigma x_i + b\Sigma x_i^2] \end{aligned} \tag{7.4}$$

These equations 7.3 and 7.4 are called normal equations. Since x_i and y_i are known, these equations result in equations in a and b. We can solve these for a and b

Remark: For the sake of simplicity leave suffix notation to obtain the following form of normal equations.

$$\Sigma y = na + b\Sigma x$$
$$\Sigma yx = a\Sigma x + b\Sigma x^2$$

EXAMPLE 7.1

Fit straight line for the following data:

x	0	2	5	7
y	-1	5	12	20

Solution: Let straight line is $a + bx$. Normal equation for straight lines are

$$\Sigma y = na + b\Sigma x$$
$$\Sigma yx = a\Sigma x + b\Sigma x^2$$

Form the following table

x	y	x^2	xy
0	-1	0	0
2	5	4	10
5	12	25	60
7	20	49	140
14	36	78	210

Now normal equations reduces to

$$36 = 4a + 14b$$
$$210 = 14a + 78b$$

On solving,

$$a = -1.1381 \qquad \text{and} \qquad b = 2.8966$$

Hence the fitted straight line is

$$y = -1.1381 + 2.8966x$$

□

■ EXAMPLE 7.2

Fit a straight line to the following data :

x	0	5	10	15	20	25
y	12	15	17	22	24	30

Solution: Here $n = 6$, i.e., even and the values of x are equally spaced , h is equal to 5. Therefore, to make the calculation easier, we take the unit of measurement $\frac{h}{2}$, i.e., 2.5 and the origin at the mean of two middle terms 10 and 15, i.e., origin is taken at $\frac{1}{2}(10+15) = 12.5$.

Taking

$$u = \frac{x - 12.5}{2.5}$$

and

$$v = y - 20$$

Let the equation of straight line be

$$v = a + bu$$

Then the normal equation are

$$\sum v = na + b\sum u$$

$$\sum uv = a\sum u + b\sum u^2$$

x	y	$u = \frac{x-12.5}{2.5}$	$v = y - 20$	uv	u^2
0	12	-5	-8	40	25
5	15	-3	-5	15	9
10	17	-1	-3	3	1
15	22	1	2	2	1
20	24	3	4	12	9
25	30	5	10	50	25
		$\sum u = 0$	$\sum v = 0$	$\sum uv = 122$	$\sum u^2 = 70$

Substituting these values in Equations (6.19) and (6.20), we have

$$0 = 6a + b0 \text{ and } 122 = a0 + 70b$$

$$\text{or } a = 0 \; b = 1.743$$

Thus, the required equation of straight line is

$$v = 0 + 1.743u \text{ or } y - 20 = 1.743\left(\frac{x-12.5}{2.5}\right)$$

$$y - 20 = \frac{1.743}{2.5}x - 1.743 \times \frac{12.5}{2.5}$$

$$\text{or } y = 0.7x + 11.285$$

□

Problems

7.1 Fit a straight line to the given data regarding x as the independent variable:

x	1	2	3	4	6	8
y	2.4	3.1	3.5	4.2	5.0	6.0

7.2 Find the best values of a and b so that $y = a + bx$ fit the given data

x	0	1	2	3	4
y	1.0	2.9	4.8	6.7	8.6

7.3 A simply supported beam carries a concentrated load $P(b)$ at its mid-point.Corresponding to various values of P. The maximum deflection $Y(in)$ is measured. The Data are given below. Find a law of the type $Y = a + bP$

P	100	120	140	160	180	200
Y	0.45	0.55	0.60	0.70	0.80	0.85

7.4 In the following table y in the weight potassium bromide which will dissolve in 100 grams of water at temperature x^0. Find a linear law between x and y.

$x^0(c)$	0	10	20	30	40	50	60	70
y gm	53.5	59.9	65.2	70.6	75.5	80.2	85.5	90

7.5 The weight of a calf taken at weekly intervals are given below. Fit a straight line using method of least squares and calculate the average rate of growth per week.

Age	1	2	3	4	5	6	7	8	9	10
weight	52.5	58.7	65	70.2	75.4	81.1	87.2	95.5	102.2	108.4

7.6 Find the least squares line for the data point $(-1,10),(0,9),(1,7),(2,5),(3,4),(4,3),(5,0)$ and $(6,-1)$.

7.7 If P is the pull required to lift a load W. By means of a pulley block, find a linear law of the firm $P = mW + C$ connectivity P and W,using the Data:

P	12	15	21	25
W	50	70	100	120

where P and W are taken in $kg - wt$.

7.1.2 Polynomials Least-Squares Fitting

The applications of the method of least squares curve fitting using polynomials are briefly discussed as follows. The least-squares mth degree Polynomials method uses mth degree polynomials

$$y = a_0 + a_1x + a_2x^2 + ... + a_mx^m \tag{7.5}$$

to approximate the given set of data $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ where $n > m$. The best fitting curve $f(x)$ has the least square error, i.e.,

$$\begin{aligned} Min\varepsilon &= Min(\Sigma_{i=1}^{n}[y_i - f(x_i)]^2) \\ &= Min(\Sigma_{i=1}^{n}[y_i - (a_0 + a_1x_i + a_2x_i^2 + ... + a_mx_i^m)]^2) \end{aligned} \tag{7.6}$$

Equating as straight line case, the first partial derivatives with respect to $a_0, a_1, ..., a_m$ to zero and simplifying, we get the following normal equations (we ignored suffix i)

$$\begin{aligned} \Sigma y &= na_0 + a_1\Sigma x + ... + a_m\Sigma x^m \\ \Sigma xy &= a_0\Sigma x + a_1\Sigma x^2 ... + a_m\Sigma x^{m+1} \\ &... \\ \Sigma x^m y &= a_0\Sigma x^m + a_1\Sigma x^{m+1} ... + a_m\Sigma x^{2m} \end{aligned}$$

These are $(n+1)$ equations in $(n+1)$ unknowns and hence can be solved for $a_0, a_1, ..., a_n$.

In particular, for fitting of a parabola $y = a + bx + cx^2$ the normal equations will be

$$\begin{aligned} \Sigma y &= na + b\Sigma x + c\Sigma x^2 \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \\ \Sigma x^2 y &= a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \end{aligned}$$

■ EXAMPLE 7.3

Fit a parabola $y = ax^2 + bx + c$ in least square sense to the data

x	10	12	15	23	20
y	14	17	23	25	21

Solution: Form the following table

	x	y	x^2	x^3	x^4	xy	x^2y
	10	14	100	1000	10000	140	1400
	12	17	144	1728	20736	204	2448
	15	23	225	3375	50625	345	5175
	23	25	529	12167	279841	575	13225
	20	21	400	8000	160000	420	8400
Sum	80	100	1398	26270	521202	1684	30648

The given numbers of data are 5. Therefore normal equations are

$$\Sigma y = a\Sigma x^2 + b\Sigma x + 5c$$
$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x$$
$$\Sigma x^2 y = a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2$$

From table, normal equations reduces to

$$100 = 1398a + 80b + 5c$$
$$1684 = 26270a + 1398b + 80c$$
$$30648 = 521202a + 26270ab + 1398c$$

On solving,

$$a = -0.07,\ b = 3.03,\ c = -8.89$$

Hence the fitted parabola is

$$y = -0.07x^2 + 3.03x - 8.89$$

□

■ EXAMPLE 7.4

Fit a second degree curve by the method of least squares for the given data :

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	1.1	1.3	1.6	2.0	2.7	3.4	4.1

Solution:

Changing the origin of x to $x - 2.5$, we get

$$u = \frac{x-2.5}{0.5} \text{ and } v = y$$

The parabolic curve is $v = a + bu + cu^2$. The normal equations are

$$\sum v = na + b\sum u + c\sum u^2$$
$$\sum uv = a\sum u + b\sum u^2 + c\sum u^3$$
$$\sum u^2 v = a\sum u^2 + b\sum u^3 + c\sum u^4$$

Here, $n = 7$

x	y	$u = \frac{x-2.5}{0.5}$	$v = y$	u^2	u^3	u^4	uv	u^2v
1.0	1.1	-3	1.1	9	-27	81	-3.3	9.9
1.5	1.3	-2	1.3	4	-8	16	-2.6	5.2
2.0	1.6	-1	1.6	1	-1	1	-1.6	1.6
2.5	2.0	0	2.0	0	0	0	0	0.0
3.0	2.7	1	2.7	1	1	1	2.7	2.7
3.5	3.4	2	3.4	4	8	16	6.8	13.6
4.0	4.1	3	4.1	9	27	81	12.3	36.9
		$\sum u = 0$	$\sum v = 16.2$	$\sum u^2 = 28$	$\sum u^3 = 0$	$\sum u^4 = 196$	$\sum uv = 14.3$	$\sum u^2 v = 69.9$

On putting the values of $\sum u, \sum u^2, \ldots$, etc. in the normal equation , we have

$$16.2 = 7a + 0b + 28c$$
$$14.3 = 0a + 28b + 0c$$
$$69.9 = 28a + 0b + 196c$$

Solving these equations, we obtain

$$a = 2.07, b = 0.511, c = 0.061$$

Hence , the curve of best fit is

$$v = 2.07 + 0.511v + 0.061v^2$$

Changing the origin , we get

$$y = 2.07 + 0.511(\frac{x-2.5}{0.5}) + 0.061(\frac{x-2.5}{0.5})^2$$

$$y = 1.04 - 0.193x - 10.243x^2$$

□

Problems

7.8 Fit a second degree parabola to the following data taking x as independent variable:

x	0	1	2	3	4
y	1	5	10	22	38

7.9 Fit a second degree parabola to the following data by least squares method:

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

7.10 The profit of certain company in the x-th year of its life are given by

x	1	2	3	4	5
y	1250	1400	1650	1950	2300

taking $u = x - 3$ and $v = y - 1650/50$, show that parabola of second degree of v on u is $v + 0.086 = 5.3u + 0.643u^2$ and deduce that the parabola of second degree of y on x is $y = 1144 + 72x + 32.15x^2$.

7.11 The Following table gives the results of the measurements of train resistances, V is the velocity in miless per hour, R is the resistance in pounds per ton:

V	20	40	60	80	100	120
R	5.5	9.1	14.9	22.8	33.3	46.0

If R is related to V by the relation $R = a + bv + cV^2$; find a, b and c by using the method of least squares.

7.2 FITTING AN EXPONENTIAL CURVE

Consider the equation $y = ae^{bx}$. Taking on both sides, we get

$$\log y = \log a + bx \log e$$

i.e.,

$$Y = A + Bx$$

where $Y = \log y$, $A = \log a$ and $B = b \log e$
This is equivalent to straight line fitting. Compute A and B from reduced normal equations

$$\Sigma Y = nA + B\Sigma x$$

$$\Sigma xY = A\Sigma x + B\Sigma x^2$$

Finally compute $a = \text{antilog } A$ and $b = \frac{B}{\log e}$

7.3 FITTING A POWER FUNCTION

Consider the equation $y = ax^b$. Taking on both sides, we get

$$\log y = \log a + b \log x$$

i.e.,

$$Y = A + bX$$

where $Y = \log y$, $A = \log a$ and $X = \log x$
This is equivalent to straight line fitting. Compute A and b from reduced normal equations

$$\Sigma Y = nA + b\Sigma X$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2$$

Finally compute $a = \text{antilog } A$

■ EXAMPLE 7.5

Find the curve of fit of the type $y = ae^{bx}$ to the following data

x	1	5	7	9	12
y	10	15	12	15	21

Solution: The curve to be fitted is $y = e^{bx}$ which is equivalent to $Y = A + Bx$, where $Y = \log y$, $A = \log a$ and $B = b \log e$
The normal equations are

$$\Sigma Y = nA + B\Sigma x$$
$$\Sigma xY = A\Sigma x + B\Sigma x^2$$

Form the following table

x	y	$Y = \log y$	x^2	xY
1	10	1.0000	1	1
5	15	1.1761	25	5.8805
7	12	1.0792	49	7.5544
9	15	1.1761	81	10.5849
12	21	1.3222	144	15.8664
$\Sigma x = 34$		$\Sigma Y = 5.7536$	$\Sigma x^2 = 300$	$\Sigma xY = 40.8862$

Substituting the values of sums, normal equations becomes

$$5.7536 = 5A + 34B$$

$$40.8862 = 34A + 300B$$

On solving,

$$A = 0.9766, B = 0.02561$$

$\therefore a = \text{antilog}A = 9.4754;\ b = \frac{B}{\log e} = 0.059$
Hence the required curve is $y = 9.4754e^{0.059x}$. □

■ EXAMPLE 7.6

Determine the curve of the form $y = ax^b$ which is best fit to the following data according to the least squares principle :

x	1.0	1.5	2.0	2.5	3.0	3.5
y	0.01	0.405	0.693	0.916	1.098	1.252

Solution:

The curve to be fitted is
$y = ax^b$
$logy = loga + blogx$
$Y = A + BX$
Then the normal equations are $\sum Y = nA + B\sum X$ and $\sum XY = A\sum X + B\sum X^2$.

x	y	$X = logx$	$Y = logy$	X^2	XY
1.0	0.01	0.0000	-2.0000	0.0000	0.0000
1.5	0.405	0.1761	-0.3925	0.3010	-0.0691
2.0	0.693	0.3010	-0.1593	0.0906	-0.0479
2.5	0.916	0.3979	-0.0381	0.1584	-0.0152
3.0	1.098	0.4771	0.0406	0.2276	0.0194
3.5	1.252	0.5441	0.0976	0.2960	0.0531
		$\sum X = 1.8963$	$\sum Y = -2.4517$	$\sum X^2 = 0.8036$	$\sum XY = -0.0597$

On putting these tabulated values in the normal equations , we have

$$-2.4517 = 6A + 1.8963B$$

$$-0.0597 = 1.8963A + 0.8036B$$

On solving , we get

$$A = -1.5149 \text{ and } B = 3.5004$$

Taking anti-log of A, we get $a = 0.0306$. Hence, the equation of curve is

$$y = 0.0306x^{3.5004}.$$

□

■ EXAMPLE 7.7

The pressure of the gas corresponding to various volumes V is measured , given by the following data :

V(cm^3)	50	60	70	90	100
P(kg/cm^2)	64.7	51.3	40.5	25.9	78

Fit the data to the equation $PV^\gamma = C$.

Solution:

We have $PV^\gamma = C$, i.e.,

$$V = C^{\frac{1}{\gamma}} P^{-\frac{1}{\gamma}}$$

Taking log of both sides , we get

$$logV = \frac{1}{\gamma} logC + (-\frac{1}{\gamma}) logP$$

$$Y = A + BX$$

where

$$Y = logV, A = \frac{1}{\gamma} logC, B = -\frac{1}{\gamma}, X = logP$$

Therefore, the normal equations are

$$\sum Y = nA + B \sum X$$

$$\sum XY = A \sum X + B \sum X^2$$

V	P	$Y = log_{10}V$	$X = log_{10}P$	XY	X^2
50	64.7	1.69897	1.81090	3.07666	3.27936
60	51.3	1.77815	1.71012	3.04085	2.92451
70	40.5	1.84510	1.60746	2.96592	2.58393
90	25.9	1.95424	1.41330	2.76193	1.99742
100	78	2.00000	1.89209	3.78418	3.58000
		$\sum Y = 9.27646$	$\sum X = 8.43387$	$\sum XY = 15.62954$	$\sum X^2 = 14.36522$

On putting these values in the normal equations , we have

$$9.27646 = 5A + 8.43387B$$

$$15.62954 = 8.43387A + 14.36522B$$

On solving , we get

$$A = 2.07042, B = -0.12754$$

$$B = -\tfrac{1}{\gamma},\ \gamma = -\frac{1}{B} = 7.84068$$

Therefore

$$\begin{aligned} A &= \frac{1}{\gamma} logC \\ \Rightarrow \log C &= A\gamma \\ &= 2.07042 \times 7.84068 = 16.23350 \\ \Rightarrow C &= 1.71199 \times 10^{16} \end{aligned}$$

Hence , the curve is

$$PV^{7.84068} = 1.71199 \times 10^{16}$$

□

7.4 FITTING A CURVE OF TYPE $Y = AX + BX^2$

Let (x_i, y_i), $i = 1, 2, ..., n$ be the n-sets of given values. Then the residual at $x = x_i$ is given by

$$E_i = y_i - ax_i - bx_i^2$$

Introducing quantity U such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=i}^{n} (y_i - ax_i - bx_i^2)^2$$

By the principle of least squares , U is minimum , when

$$\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b}$$

Therefore

$$\frac{\partial U}{\partial a} = \sum_{i=1}^{n} 2(y_i - ax_i - bx_i^2)(-x_i) = 0$$

$$\sum xy = a\sum x^2 + b\sum x^3 \tag{7.7}$$

and

$$\frac{\partial U}{\partial b} = \sum_{i=1}^{n} 2(y_i - ax_i - bx_i^2)(-x_i^2) = 0$$

$$\sum x^2 y = a\sum x^3 + b\sum x^4 \tag{7.8}$$

Equations 7.7 and 7.8 are normal equations. Solving Equations 7.7and 7.8, we get a and b.

7.5 FITTING A CURVE OF TYPE $XY = A + BX$

Let the given n points be $(x_1, y_1), (x_2, y_2).....(x_n, y_n)$.Rearranging the given equation , we have

$$y = \frac{a}{x} + b$$

Then the residual at $x = x_i$is

$$E_i = (y_i - \frac{a}{x_i} - b)$$

Introducing quantity u such a way that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - \frac{a}{x_i} - b)^2$$

By the principle of least squares , we have

$$\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b}$$

$$\therefore \frac{\partial U}{\partial a} = \sum_{i=1}^{n} 2(y_i - \frac{a}{x_i} - b)(-\frac{1}{x_i}) = 0$$

$$\sum \frac{y}{x} = a \sum \frac{1}{x^2} + b \sum \frac{1}{x} \tag{7.9}$$

$$\frac{\partial U}{\partial b} = \sum_{i=1}^{n} 2(y_i - \frac{a}{x_i} - b)(-1) = 0$$

$$\sum y = a \sum \frac{1}{x} + b_n \tag{7.10}$$

Equations 7.9 and 7.10 are normal equations. On solving these equations, we obtain the values of a and b.

7.6 FITTING A CURVE OF TYPE $Y = AX + \frac{B}{X}$

Let the given n points be $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$. Then the residual for $x = x_i$ is

$$E_i = y_i - ax_i - \frac{b}{x_i}$$

and

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - ax_i - \frac{b}{x_i})^2$$

By the principle of least squares , we have

$$\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b}$$

$$\text{i.e., } \frac{\partial U}{\partial a} = \sum_{i=1}^{n} 2(y_i - ax_i - \frac{b}{x_i})(-x_i) = 0$$

$$\sum xy = a \sum x^2 + nb \tag{7.11}$$

$$\text{and } \frac{\partial U}{\partial b} = \sum_{i=1}^{n} 2(y_i - ax_i - \frac{b}{x_i})(-\frac{1}{x_i}) = 0$$

$$\sum \frac{y}{x} = na + b \sum \frac{1}{x^2} \tag{7.12}$$

Equations 7.11 and 7.12 are normal equations. On solving them , we obtain the values of a and b.

7.7 FITTING A CURVE OF TYPE $Y = A + \frac{B}{X} + \frac{C}{X^2}$

Let the given points be $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$. Then the residual for $x = x_i$ is given by

$$E_i = y_i - a - \frac{b}{x_i} - \frac{c}{x_i^2}$$

and

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - a - \frac{b}{x_i} - \frac{c}{x_i^2})^2$$

By the principle of least squares , we have

$$\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b} = \frac{\partial U}{\partial c}$$

$$\frac{\partial U}{\partial a} = \sum_{i=1}^{2} 2(y_i - a - \frac{b}{x_i} - \frac{c}{x_i^2})(-1) = 0$$

$$\sum y = na + b\sum \frac{1}{x} + c\sum \frac{1}{x^2} \tag{7.13}$$

$$\frac{\partial U}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - \frac{b}{x_i} - \frac{c}{x_i^2})(-\frac{1}{x_i}) = 0$$

$$\sum \frac{y}{x} = a\sum \frac{1}{x} + b\sum \frac{1}{x^2} + c\sum \frac{1}{x^3} \tag{7.14}$$

and $\frac{\partial U}{\partial c} = \sum_{i=1}^{n} 2(y_i - a - \frac{b}{x_i} - \frac{c}{x_i^2})(-\frac{1}{x_i^2}) = 0$

$$\sum \frac{y}{x^2} = a\sum \frac{1}{x^2} + b\sum \frac{1}{x^3} + c\sum \frac{1}{x^4} \tag{7.15}$$

Equations 7.13 , 7.14 , and 7.15 are normal equations. On solving these equations, we obtain the values of a, b, and c.

7.8 FITTING A CURVE OF TYPE $Y = \frac{A}{X} + B\sqrt{X}$

Let the given points be $(x_1, y_1), (x_2, y_2) \ldots\ldots, (x_n, y_n)$. Then the residual for $x = x_i$is given by

$$E_i = y_i - \frac{a}{x_i} - b\sqrt{x_i}$$

Introducing quantity U such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - \frac{a}{x_i} - b\sqrt{x_i})^2$$

By the principle of least squares , we have U minimum

$$\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b}$$

$$\frac{\partial U}{\partial a} = \sum_{i=1}^{n} 2(y_i - \frac{a}{x_i} - b\sqrt{x_i})(-\frac{1}{x_i}) = 0$$

$$\sum \frac{y}{x} = a\sum \frac{1}{x^2} + b\frac{1}{\sqrt{x}} \tag{7.16}$$

and

$$\frac{\partial U}{\partial b} = \sum_{i=1}^{n} 2(y_i - \frac{a}{x_i} - b\sqrt{x_i})(-\sqrt{x_i}) = 0$$

or

$$\sum \sqrt{xy} = a\sum \frac{1}{\sqrt{x}} + b\sum x \tag{7.17}$$

Equation 7.16 and 7.17 are normal equations to the curve $y = \frac{a}{x} + b\sqrt{x}$.On solving these equations , one can obtain the values of a and b.

7.9 FITTING A CURVE OF TYPE $2^X = AX^2 + BX + C$

By the principle of least squares , one can obtain the normal equations to the curve as follows:

$$\sum 2^x x^2 = a\sum x^4 + b\sum x^3 + c\sum x^2$$

$$\sum 2^x x = a\sum x^3 + b\sum x^2 + c\sum x$$

$$\text{and } \sum 2^x = a\sum x^2 + b\sum x + nc$$

where n is the number of points (x_i, y_i), $0 \le i \le n$.

EXAMPLE 7.8

Fit a curve $y = ax + \frac{b}{x}$ to the data given below :

x	1	2	3	4	5	6	7	8
y	5.43	6.28	8.23	10.32	12.63	14.84	17.27	19.51

Solution:

The curve to be fitted is

$$y = ax + \frac{b}{x}$$

Then the normal equations are

$$\sum xy = a\sum x^2 + nb$$

$$\frac{\sum y}{x} = na + b\sum \frac{1}{x^2}$$

Here $n = 8$.

x	y	x^2	xy	$\frac{y}{x}$	$\frac{I}{x^2}$
1	5.43	1	5.43	5.43	1
2	6.28	4	12.56	3.14	0.25
3	8.23	9	24.69	2.7433	0.1111
4	10.32	16	41.28	2.5800	0.0625
5	12.63	25	63.15	2.5260	0.0400
6	14.84	36	89.04	2.4733	0.0278
7	17.27	49	120.89	2.4671	0.0204
8	19.51	64	156.08	2.4388	0.0156
		$\sum x^2 = 204$	$\sum xy = 513.12$	$\sum \frac{y}{x} = 23.7985$	$\sum \frac{1}{x^2} = 1.5274$

On putting these values of $\sum xy$, $\frac{\sum y}{x}$, ...*etc*. in the normal equations, we get

$$513.12 = 204a + 8b$$

$$23.7985 = 8a + 1.5274b$$

On solving these equations , we get

$$b = 3.0289, a = 2.3965$$

Hence , the equation of the curve is

$$y = 2.3965x + \frac{3.0289}{x}$$

□

■ EXAMPLE 7.9

Fit a curve of the type $xy = ax + b$ to the following data :

x	1	3	5	7	9	10
y	36	29	28	26	24	15

Solution:

The equation of the given curve is

$$xy = ax + b$$

$$y = a + \frac{b}{x}$$

Therefore , the normal equations are

$$\sum y = na + b\sum \frac{1}{x}$$

$$\frac{\sum y}{x} = a\sum \frac{1}{x} + b\sum \frac{1}{x^2}$$

Here $n = 6$.

x	y	$\frac{1}{x}$	$\frac{1}{x^2}$	$\frac{y}{x}$
1	36	1	1	36
3	29	0.3333	0.1111	9.6667
5	28	0.2000	0.4000	5.6000
7	26	0.1429	0.0204	3.7143
9	24	0.1111	0.0123	2.6667
10	15	0.1000	0.0100	1.5
	$\sum y = 158$	$\sum \frac{1}{x} = 1.8873$	$\sum \frac{1}{x^2} = 1.1938$	$\sum \frac{y}{x} = 59.1477$

On putting these values in the above normal equations, we have

$$158 = 6a + 1.8873b$$

$$59.1477 = 1.8873a + 1.1938b$$

On solving the equations, we get

$$a = 21.3810, b = 15.7441$$

Hence, the equation of the curve is

$$xy = 21.3810x + 15.7441$$

□

EXAMPLE 7.10

A person runs the same racetrack for five consecutive days and is timed as follows:

Days(x)	1	2	3	4	5
Time(y)	15.3	15.1	15	14.5	14

Make a least squares fit to the above data using the function $y = a + \frac{b}{x} + \frac{c}{x^2}$.

Solution: The given equation is

$$y = a + \frac{b}{x} + \frac{c}{x^2}$$

Therefore , the normal equations are

$$\sum y = na + b\sum \frac{1}{x} + c\sum \frac{1}{x^2}$$

$$\frac{\sum y}{x} = a\sum \frac{1}{x} + b\sum \frac{1}{x^2} + c\sum \frac{1}{x^3}$$

$$\frac{\sum y}{x^2} = a\sum \frac{1}{x^2} + b\sum \frac{1}{x^3} + c\sum \frac{1}{x^4}$$

Here n = 5.

x	y	$\frac{1}{x}$	$\frac{1}{x^2}$	$\frac{1}{x^3}$	$\frac{1}{x^4}$	$\frac{y}{x}$	$\frac{y}{x^2}$
1	15.3	1	1	1	1	15.3	15.3
2	15.1	0.5	0.22	0.1250	0.0625	7.5500	3.7550
3	15	0.3333	0.1111	0.0370	0.0123	5.0000	1.6667
4	14.5	0.2500	0.0625	0.0156	0.0039	3.6250	0.9063
5	14	0.2000	0.0400	0.0080	0.0016	2.8000	0.5600
	$\sum y = 73.9$	$\sum \frac{1}{x} = 2.2833$	$\sum \frac{1}{x^2} = 1.4636$	$\sum \frac{1}{x^3} = 1.1856$	$\sum \frac{1}{x^4} = 1.0803$	$\sum \frac{y}{x} = 34.2750$	$\sum \frac{y}{x^2} = 22.1880$

On putting the values $\sum y, \sum \frac{1}{x}, ...etc$,in the normal equations, we get

$$73.9 = 5a + 2.283b + 1.4636c$$

$$34.2750 = 2.2833a + 1.4636b + 1.1856c$$

$$22.1880 = 1.4636a + 1.1856b + 1.0803c$$

On solving these equation , we get

$$a = 12.6751, b = 8.2676, \text{ and } c = -5.7071$$

Hence , the equation is

$$y = 12.6751 + \frac{8.2676}{x} - \frac{5.7071}{x^2}$$

□

■ **EXAMPLE 7.11**

Use the method of least squares to fit the curve $y = \frac{c_0}{x} + c_1\sqrt{x}$ to the following table of values :

x	0.1	0.2	0.4	0.5	1	2
y	21	11	7	6	5	6

Solution:
The given equation is $y = \frac{c_0}{x} + c_1\sqrt{x}$, the normal equations are

$$\sum \frac{y}{x} = c_0 \sum \frac{1}{x^2} + c_1 \sum \frac{1}{\sqrt{x}}; \ \sum \sqrt{xy} = c_0 \sum \frac{1}{\sqrt{x}} + c_1 \sum x$$

x	y	$\frac{y}{x}$	$\sqrt{xy}$	$\frac{1}{\sqrt{x}}$	$\frac{1}{x^2}$
0.1	21	210	6.6408	3.1623	100
0.2	11	55	4.9193	2.2361	25
0.4	7	17.5	4.4272	1.5811	6.25
0.5	6	12	4.2426	1.4142	4.00
1	5	5	5	1.0000	1
2	6	3	8.4853	0.7071	0.25
$\sum X = 4.2$		$\sum \frac{y}{x} = 302.5$	$\sum \sqrt{xy} = 33.7152$	$\sum \frac{1}{\sqrt{x}} = 10.1008$	$\sum \frac{1}{x^2} = 136.50$

On putting these values in the normal equations , we have

$$302.5 = 136.5c_0 + 10.1008c_1$$

$$33.7152 = 10.1008c_0 + 4.2c_1$$

On solving , we get $c_0 = 1.9733, c_1 = 3.2818$ Hence the equation of the curve is,

$$y = \frac{1.9733}{x} + 3.2818\sqrt{x}$$

□

Problems

7.12 Fit a equation of the form $y = ae^{bx}$ to the following data by the method of least squares.

x	1	2	3	4
y	1.65	2.7	4.5	7.35

7.13 The voltage V across a capacitor at time t seconds is given by the following table. Use the principle of least squares to fit a curve of the form $V = ae^{bt}$ to the data

t	0	2	3	4	6	8
v	150	63	22	28	12	5.6

7.14 Fit a curve of the form $y = ae^{bx}$ to the following data:

x	0	2	4
y	5.102	10	31.48

7.15 Fit a curve of the $formy = ax^b$ to the data given below:

x	1	2	3	4	5
y	7.1	27.8	62.1	110	161

7.16 Fit a curve of the form $y = ab^x$ in least square sense to the data given below:

x	2	3	4	5	6
y	144	172.8	207.4	248.8	298.5

7.17 Drive the least square equations for fitting a curve of the type $y = ax^2 + \frac{b}{x}$ to set of n points. Hence fit a curve of this type to the data:

x	1	2	3	4
y	-1.51	0.99	3.88	7.66

7.18 Drive the least squares approximation of the type $ax^2 + bx + c$ to the function 2^x at the point $X_i = 0, 1, 2, 3, 4$.

7.19 A person runs the same race track for 5 consecutive days and is timed as follows:

day x	1	2	3	4	5
day y	15.3	15.1	15	14.5	14

Make a least square fit to the above data using a function $a + \frac{b}{x} + \frac{c}{x^2}$.

7.20 It is known that the variable x and y hold the relation of the form $y = ax + \frac{b}{x}$. Fit the curve to the given data:

x	1	2	3	4	5	6	7	8
y	5.43	6.28	8.23	10.32	12.63	14.86	17.27	19.51

7.21 Fit a carve of the type $xy = ax + b$ to the following data:

x	1	3	5	7	9	10
y	36	29	28	26	24	15

7.22 Determine the contant of the curve $y = ax + bx^2$ for the following data:

x	0	1	2	3	4
y	2.1	2.4	2.6	2.7	3.4

7.23 The pressure and volume of a gas are related by the equation $pv^\gamma = b$ where a and b are constant. Fit this equation to the following set of data:

p(kg/cm^3)	0.5	1	1.5	2	2.5	3
v(liters)	1.62	1	0.75	0.62	0.52	0.46

CHAPTER 8

CORRELATION

When for every value of a variable X we known a corresponding value of a second variable Y, i.e., the data is in the form of paired measurements,then we are interested in the relationship of these two variables.

The analysis of univariate data in case of economic, social and scientific areas becomes insufficient. So, in some situations, say production price, height of father and son , marks obtained in two subjects, height of husbands and their wives, a series of expenses on advertisement and sales.

If change in one variable is accompanied(or appears to be accompanied) a change in other variable and vice-versa, then the two variables are said to be correlated and this relationship is known as correlation or covariation.

8.1 CORRELATION ANALYSIS

The methods that are employed to determine if there exists any relationship between two variables and to express this relationship numerically comes under correlation analysis. Correlation analysis was developed by Francis Galton and Karl Pearson. Here we should consider only a logical relationship.

Nature of the Relationship or Factors Responsible for Relationship

1. Direct Relationship : One variable may be the cause of the other. As to which is the cause and which the effect is to be judged from the circumstances.

2. Common Cause : Correlation may be due to any other common cause. That is both variables may be the result of a common cause. For example, the heights of parents and their offspring are related due to their blood relation.

3. Mutual Reaction: It is not necesary that a series will affect the other.The two series may affect each other. Then it is not possible to differentiate between cause and effect.
4. Useless or Nonsense or Spurious Correlation : It might sometimes happen that between two variables a relationship may be observed when none of such relationship exists in the universe. Such a correlation is known as Spurious correlation. For example, the relationship between the number of tourists and production of sugar in India.

This means, the presence of correlation between two variables X and Y does not necessarily imply the existence of direct causation, though causation will always result in correlation.

Spurious correlation practically has no meaning and we may consider that the observed correlation is due to chance.

Utility and Importance of Correlation Analysis. Correlation analysis is a very important technique in statistics. It is useful in physical and social sciences and business and economics.

The correlation analysis is useful in the following cases :

1. To have more reliable forecasting.
2. To study economic activities.
3. To estimate the variable values on the basis of an other variable values.
4. To make analysis, drawing conclusions etc. in the research or statistical investigations.

In economics, we study the relationship between price and demand, price and supply, income and expenditure, etc. According to Neiswanger,

> Correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variables on which others depend . may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing force may become effective.

To a businessman correlation analysis helps to estimate costs , sales, prices and other related variables.

Correlation analysis is the basis of the concept of regression and ratio of variation.

According to tippet,

> The effects of the correlation is to reduce the range of uncertainty of our prediction.

8.2 TYPES OF CORRELATION

On the basis of direction correlation.

1. Positive correlation and
2. Negative correlation.

When the increase (or decrease) in one variable results in a corresponding increase (or decrease) in the other (i.e., the two variables tend to move together in the same direction), the correlation is said to be positive or direct correlation. For example, demand and price.

When the increase (or decrease) in one variable results in a corresponding decrease (or increase) in the other (i.e., the two variable tend to move together in opposite direction) , the correlation said to be negative or indirect correlation. For example , supply and price.

On the basis of ratio of change .

1. Linear correlation
2. Non-linear correlation.

If the ratio of change between the two variables is uniform i.e., when the amount of movement in one variable bears a constant ratio to the amount of change in the order, the correlation is said to be linear otherwise it is called non- linear or curvilinear.

On the basis of the number of variables.

1. Simple correlation ,
2. Multiple correlation and
3. partial correlation.

The correlation between two variables is called simple correlation, when the number of variables is more than two we study multiple correlation and partial correlation.

8.3 METHODS OF DETERMINING CORRELATION

There are various methods of determining correlation in two variables. Here we discuss only following methods:
(1) Karl Pearson's Coefficient of correlation
(2) Spearman's rank Correlation method
(3) Method of Least Squares.

8.4 KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl Pearson, a great biometrician and statistician suggested in 1890, a mathematical method for measuring the magnitude of relationship between two variables. It gives the most widely used formula, called Pearson's coefficient of correlation and denoted by r. This formula is also called product moment correlation coefficient.

Karl Pearson's Correlation Coefficient,

$$r = \frac{\text{Covariance between } X \text{ and } Y}{(\text{Standard Deviation of } X)(\text{Standard Deviation of } Y)}$$

That is

$$r = \frac{\text{Cov}(X,Y)}{\sigma_x \times \sigma_y}$$

where,

$$\text{Cov}(x,y) = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{N} = \frac{\sum d_x d_y}{N}$$

$$\sigma_x = \sqrt{\frac{\sum(X-\overline{X})^2}{N}} or \sqrt{\frac{\sum d_x^2}{N}}$$

$$\sigma_y = \sqrt{\frac{\sum(Y-\overline{Y})^2}{N}} or \sqrt{\frac{\sum d_y^2}{N}}$$

where, $d_x = X - \overline{X}$ =deviation of series X from its mean $d_y = Y - \overline{X}$ =deviation of series Y from its mean
Thus,

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2 \sum(Y-\overline{Y})^2}}$$

8.4.1 Main Characteristics of Karl Pearson's Coefficient of Correlation

1. It is an ideal measure of correlation and is independent of the units of X and Y.
2. It is independent of change of origin and scale.
3. It is based on all the observations.

4. It various between -1 and $+1$:
   (a) $r = -1$, when there is a perfect negative correlation
   (b) $r = 0$, when there is no correlation
   (c) $r = +1$, when there is a perfect positive correlation.
5. It does not tell any thing about cause and effect relationship.
6. It is some how difficult to calculate.
7. It requires some interpretation.

8.4.2 Assumptions of Karl Pearson's Coefficient of Correlation

The Karl Pearson's coefficient of correlation is based on the following three assumptions :

1. In each of the series a large number of independent causes are operating so as to produce normal distribution.
2. The forces so operated has a relationship of cause and effect.
3. The relationship between two variables is linear.

8.4.3 Degree of Correlation

The relationship between two variables degrees as explained below :

1. Perfect Correlation : The correlation is said to be be perfect when a change in one variable is always followed by a corresponding proportional change in the other. If this change is in the same direction it is said to be perfect positive correlation and if this change is in the opposite direction it is said to be perfect negative correlation .
2. Absence of Correlation : When a change in one variable does not effect another variable, then we call it no correlation.
3. Limited Degrees of correlation : When the change in one variable is always followed by a corresponding change in other variable but not in the same proportion, then correlation is said to be limited. Limited degrees of correlation are :
   (a) High Degree of Correlation
   (b) Moderate Degree of Correlation

When correlation is measured with help of Karl Pearson's coefficient of correlation,r, then we have the results :

Correlation	Positive Value of r	Negative Value of r
Perfect Correlation	$+1$	-1
Very High Degree Correlation	between .9 to .99	between -0.99 to -0.9
High Degree Correlation	between .75 and .9	between-0.9 to -0.75
Moderate Degree Correlation	between .25 and .75	between -0.75 and -0.25
Low Degree Correlation	between 0 and .25	between -0.25 and 0
No Correlation	0	0

8.4.4 Methods of Calculation of Karl Pearson's Coefficient of Correlation

8.4.4.1 Actual Mean Method for Computing Correlation Coefficient

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2\sum(Y-\overline{Y})^2}}$$

or

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}$$

where $d_x = X - \overline{X}, d_y = Y - \overline{Y}$

Steps:.

1. Find the means of the two series i.e., find $\overline{X}$ and $\overline{Y}$.
2. Take deviations of the two series from $\overline{X}$ and $\overline{Y}$ respectively i.e., calculate d_x and d_y.
3. Square these deviations and get the totals , i.e.,find $\sum d_x^2$ and $\sum d_y^2$.
4. Multiply the deviations $d_x$and $d_y$and get the total i.e., find $\sum d_x d_y$.
5. Calculate $\sigma_x$and $\sigma_y$the respective standard deviations of the two series.
6. Substitute these value in the formula :

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}$$

■ EXAMPLE 8.1

Calculate Karl Pearson's correlation coefficient from the following data :

X	11	10	9	8	7	6	5
Y	20	18	12	8	10	5	4

Solution:

X	$d_x = X - \overline{X}$	d_x^2	Y	$d_y = Y - \overline{Y}$	d_y^2	$d_x d_y$
11	3	9	20	9	81	27
10	2	4	18	7	49	14
9	1	1	12	1	1	1
8	0	0	8	-3	9	0
7	-1	1	10	-1	1	1
6	-2	4	5	-6	36	12
5	-3	9	4	-7	49	21
$\sum X = 56$	0	$\sum d_x^2 = 28$	$\sum Y = 77$	0	$\sum d_y^2 = 226$	$\sum d_x d_y = 76$

$$\overline{X} = \frac{\sum X}{N} = \frac{56}{7} = 8$$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{77}{7} = 11$$

Karl Pearson's Coefficient of Correlation,

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \times \sum d_y^2}} = \frac{76}{\sqrt{28 \times 226}} = \frac{76}{79.55} = 0.96$$

EXAMPLE 8.2

Calculate coefficient of correlation between the marks obtained by 10 students in Accountancy and statistics:

Student	1	2	3	4	5	6	7	8	9	10
Accountancy	45	70	65	30	90	40	50	75	85	60
Statistics	35	90	70	40	95	40	60	80	80	50

Solution:

Student	X	$d_x = X - \bar{X}$	d_x^2	Y	$d_y = Y - \bar{Y}$	d_y^2	$d_x d_y$
1	45	-16	256	35	-29	841	464
2	70	9	81	90	26	676	234
3	65	4	16	70	6	36	24
4	30	-31	961	40	-24	576	744
5	90	29	841	95	31	961	899
6	40	-21	441	40	-24	576	504
7	50	-11	121	60	4	16	44
8	75	14	196	80	16	256	224
9	85	24	576	80	16	256	384
10	60	-1	1	50	-14	196	14
N = 10	610	0	3490	640	0	4390	3535

$$\bar{X} = \frac{\sum X}{N} = \frac{610}{10} = 61$$
$$\bar{Y} = \frac{\sum Y}{N} = \frac{640}{10} = 64$$
$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \times \sum d_y^2}} = \frac{3535}{\sqrt{3490 \times 4390}} = 0.903$$

□

EXAMPLE 8.3

Compute Karl Pearson's coefficient of correlation from the following data taking deviations from actual means $\bar{X} = 66$ and $\bar{Y} = 65$:

X	84	51	91	60	68	62	86	58	53	47
Y	78	36	98	25	?	82	90	62	65	39

Solution: Since one value in Y-series is not known, we use the relation $\sum Y = N\bar{Y}$ to obtain this value. Let the unknown value be a, Then

$$\sum Y = N\bar{Y}$$
$$575 + a = 10 \times 65$$
$$a = 75$$

Computing Table for Correlation Coefficient,

X	$d_x = X - 66$	d_x^2	Y	$d_x = Y - 66$	d_y^2	$d_x d_y$
84	18	324	78	13	169	234
51	-15	225	36	-29	841	435
91	25	625	98	33	1089	825
60	-6	36	25	-40	1600	240
68	2	4	$a = 75$	10	100	20
62	-4	16	82	17	289	-68
86	20	400	90	25	625	500
58	-8	64	62	-3	9	24
53	-13	169	65	0	0	0
47	-19	361	39	-26	676	494
660	_	2224	775+75=650	_	5398	2704

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \times d_y^2}}$$

$\therefore$

$$r = \frac{2704}{\sqrt{2224 \times 5398}} \approx 0.78$$

□

■ EXAMPLE 8.4

From the following table, determine $\sigma_x, \sigma_y$and the coefficient of correlation between X and Y series :

	X -series	Y-series
No. of Items	15	15
Arithmetic Mean	25	18
Sum of Squares of Deviation from Mean	136	138

Sum of products of deviations of X and Y-Series from mean =122.

Solution: Given: $N = 15, \overline{X} = 25, \overline{Y} = 18, \sum d_x^2 = 136, \sum d_y^2 = 138, \sum d_x d_y = 122$

$$\sigma_x = \sqrt{\frac{\sum d_x^2}{n}} = \sqrt{\frac{136}{15}} = 3.01$$

$$\sigma_y = \sqrt{\frac{\sum d_y^2}{n}} = \sqrt{\frac{138}{15}} = 3.03$$

Karl Pearson's Correlation Coefficient

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 . \sum d_y^2}} = \frac{122}{\sqrt{136 \times 138}} \approx 0.89$$

□

8.4.4.2 Shortcut Method for Computing Correlation Coefficient *When actual mean is not a whole number*, then we use short-cut method in which deviations are taken from assumed mean.

Steps:.

1. Take the deviation of X-series from an assumed mean A_x, say and denote these deviations by $d_x$and get the total $\sum d_x$.
2. Take the deviation of Y-series from an assumed mean A_y,say and denote these deviations by $d_y$and get the total $\sum d_y$.
3. Square d_x and get the total $\sum d_x^2$.
4. Square d_y and get the total $\sum d_y^2$.
5. Multiply $d_x$and $d_y$and get the total $\sum d_x d_y$.
6. Use the formula given below in any one of the forms

$$r = \frac{N\sum d_x d_y - \sum d_x . \sum d_y}{\sqrt{[N\sum d_x^2 - (\sum d_x)^2][N\sum d_y^2 - (\sum d_y)^2]}}$$

where N = Number of pairs, A_x= Assumed mean of X-series, A_y= Assumed mean of Y-series, $\sum d_x = X - A_x$= Sum of the deviations from assumed mean in X-series, $\sum d_y = Y - A_y$=Sum of the deviations from assumed mean in Y-series, $\sum d_x d_y$= Sum of the product of deviations from assumed means, $\sum d_x^2$ =Sum of the squares of the deviations from assumed mean in X-series, $\sum d_y^2$ =Sum of the squares of the deviations from assumed mean in Y-series, σ_x = Standard deviation of X-series, σ_y = Standard deviation of Y-series.

■ EXAMPLE 8.5

Find the coefficient of correlation between the values of X and Y given below :

X	78	89	97	69	54	79	60	65
Y	125	137	156	112	107	136	120	110

Solution:

X	$d_x = X - 70$	d_x^2	Y	$d_y = Y - 120$	d_y^2	$d_x d_y$
78	8	49	125	5	25	40
89	19	361	137	17	289	323
97	27	729	156	36	1296	972
69	-1	1	112	-8	64	8
54	-16	256	107	-13	169	208
79	9	81	136	16	256	144
60	-10	100	120	0	0	0
65	-5	25	110	-10	100	50
-	$\sum d_x = 31$	$\sum d_x^2 = 1602$	-	$\sum d_y = 43$	$\sum d_y^2 = 2199$	$\sum d_x d_y = 1745$

$$\begin{aligned} r &= \frac{N\sum d_x d_y - \sum d_x . \sum d_y}{\sqrt{\{N\sum d_x^2 - (\sum d_x)^2\}}\sqrt{\{N\sum d_y^2 - (\sum d_y)^2\}}} \\ &= \frac{8 \times 1745 - 31 \times 43}{\sqrt{\{8 \times 1602 - (31)^2\}}\sqrt{\{8 \times 2199 - (43)^2\}}} \\ &= \frac{13960 - 1333}{\sqrt{\{12816 - 961\}}\sqrt{\{17592 - 1849\}}} \\ &= 0.924 \end{aligned}$$

□

8.4.4.3 Step Deviation Method for Computing Correlation Coefficient Let the two variables be X and Y . Let

$$U = \frac{X-a}{h} \text{ and } V = \frac{Y-b}{k}$$

Then the correlation coefficient between U and V and correlation coefficient between X and Y are equal . That is

$$r_{xy} = r_{uv}$$

This implies correlation coefficient is independent of the change of origin and scale.

However if,

$$U = \frac{X-a}{h} \text{ and } V = \frac{Y-b}{-k}; h > 0, k > 0$$

Then

$$r_{xy} = -r_{uv}$$

The procedure is same as explained in shortcut method. If, we take $d_x = U$ and $d_y = V$, the formula is again

$$r = \frac{N\sum d_x d_y - \sum d_x . \sum d_y}{\sqrt{[N\sum d_x^2 - (\sum d_x)^2][N\sum d_y^2 - (\sum d_y)^2]}}$$

where symbols have their usual meaning.

■ **EXAMPLE 8.6**

Total sales turnover and net profit of seven medium sized companies are given below. Calculate the Karl Pearson's correlation coefficient:

Sales turnover in ton	100	200	300	400	500	600	700
Net Profit ('000 Rs.)	30	50	60	80	100	110	130

Solution: Let $d_x = U = \frac{X-400}{100}$ and $d_y = V = \frac{Y-80}{10}$. Then $r_{xy} = r_{uv}$

X	d_x	d_x^2	Y	d_y	d_y^2	$d_x d_y$
100	-3	9	30	-5	25	15
200	-2	4	50	-3	9	6
300	-1	1	60	-2	4	2
400	0	0	80	0	0	0
500	1	1	100	2	4	2
600	2	4	110	3	9	6
700	3	9	130	5	25	15
Total	$\sum d_x = 0$	$\sum d_x^2 = 28$	-	$\sum d_y = 0$	$\sum d_y^2 = 76$	$\sum d_x d_y = 46$

Here, $N = 7$

$$r = \frac{N\sum d_x d_y - \sum d_x . \sum d_y}{\sqrt{[N\sum d_x^2 - (\sum d_x)^2][N\sum d_y^2 - (\sum d_y)^2]}}$$

$$= \frac{7 \times 46 - 0}{\sqrt{\{7 \times 28 - 0\}}\sqrt{7 \times 76 - 0}} \approx 0.997$$

□

8.4.4.4 Direct Method: Computation of Coefficient of Correlation By Square of Values
Use this method if assumed mean is zero.

Steps:.

1. Calculate total for each series i.e., find $\sum X$and $\sum Y$.
2. Find the square of all values in X-series and get $\sum X^2$
3. Find the square of all values in Y-series and get $\sum Y^2$.
4. Find the product of X and Y and get $\sum XY$.
5. Substitute the values in the formula and simplify.

Formula

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2} \times \sqrt{N\sum Y^2 - (\sum Y)^2}}$$

EXAMPLE 8.7

Find the coefficient of correlation between X and Y variables for the data given below:

X	17	18	19	19	20	20	21	21	22	23
Y	12	16	14	11	15	19	22	16	15	20

Solution:

X	X^2	Y	Y^2	XY
17	289	12	144	204
18	324	16	256	288
19	361	14	196	266
19	361	11	121	209
20	400	15	225	300
20	400	19	361	380
21	441	22	484	462
21	441	16	256	336
22	484	15	225	330
23	529	20	400	460
200	4030	160	2668	3235

$$\begin{aligned} r &= \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2} \times \sqrt{N\sum Y^2 - (\sum Y)^2}} \\ &= \frac{10 \times 3235 - 200 \times 160}{\sqrt{10 \times 4030 - (200)^2} \times \sqrt{10 \times 2668 - (160)^2}} \\ &= 0.615 \end{aligned}$$

□

EXAMPLE 8.8

Calculate the number of items for which $r = +0.8$, $\sum xy = 200$, standard deviation of $y = 5$ and $\sum x^2 = 100$,where x and y denote deviation of items from actual mean (where $\sum x = 0 = \sum y$).

Solution:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

$$
\begin{aligned}
&\Rightarrow 0.8 = \frac{200}{\sqrt{100 \times \sum y^2}} \\
&\Rightarrow 0.64 \times 100 \sum y^2 = 40000 \\
&\Rightarrow \sum y^2 = \frac{40000}{64} = 625 \\
&\because \sigma_y = \sqrt{\frac{\sum y^2}{N}} \\
&\therefore 5 = \sqrt{\frac{625}{N}} \text{Given } \sigma_y = 5 \\
&\Rightarrow N = 25
\end{aligned}
$$

□

■ EXAMPLE 8.9

If covariance between x and y variables is 9.6 and the variance of x and y are respectively 16 and 9 , find the coefficient of correlation.

Solution: Given,

$$Cov(x,y) = \frac{\sum xy}{N} = 9.6$$

Variance of x,

$$\sigma_x^2 = 16 \;\Rightarrow\; \sigma_x = 4$$

Variance of y

$$\sigma_y^2 = 9 \;\Rightarrow\; \sigma_y = 3$$

$$\therefore r = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} = \frac{9.6}{4 \times 3} = 0.8$$

□

■ EXAMPLE 8.10

A student while computing the correlation coefficient between two variables X and Y obtained the following results:

$$N = 25, \sum X = 125, \sum X^2 = 650, \sum Y = 100, \sum Y^2 = 460, \sum XY = 508.$$

It was, however, discovered later that he had copied down to pairs as observations as (6, 14),(8, 6), while the correct values were (8, 12), (6, 8). Correct this error and find the Karl Pearson's correlation coefficient.

Solution: Subtract wrong figures from each total and add correct figure to get correct totals.

$$
\begin{aligned}
\text{Correct} \sum X &= 125 - (6+8) + (8+6) = 125 \\
\text{Correct} \sum X^2 &= 650 - (6^2+8^2) + (8^2+6^2) = 650 \\
\text{Correct} \sum Y &= 100 - (14+6) + (12+8) = 100 \\
\text{Correct} \sum Y^2 &= 460 - (14^2+6^2) + (12^2 + +8^2) = 436 \\
\text{Correct} \sum XY &= 508 - (6 \times 14 + 8 \times 6) + (8 \times 12 + 6 \times 8) = 520
\end{aligned}
$$

Hence coerrect r is given by

Problems

8.1 Calculate Karl Pearson's correlation coefficient r from the following data:

a)

X	40	44	42	43	44	45
Y	56	54	60	64	62	58

b)

X	20	25	30	35	40	45	50	55	60
Y	16	20	23	25	33	38	46	50	55

8.2 The sale and expenditure of 10 companies is given below, compute the coefficient of correlation:

Company	1	2	3	4	5	6	7	8	9	10
Sales ('00000)	50	50	55	60	65	60	65	60	60	50
Expenditure ('000)	11	13	14	16	16	15	15	14	13	13

8.3 Calculate Karl Pearson's coefficient of correlation between the ages of husbands and wives

Husband's Age	23	27	28	29	30	31	33	35	36	39
Wife's Age	18	22	23	24	25	25	28	29	30	32

8.4 A study is made relating aptitude scores (X) to productivity index (Y) in a factory after six months training of personnel. The following are the figures regarding seven randomly selected workers:

X	10	20	30	40	50	60	70
Y	3	5	6	8	10	11	13

Find correlation coefficient between X and Y.

8.5 Calculate Karl Pearson's coefficient of correlation between advertisement expenditure (X) and profit (Y) for 9 months:

X	300	350	400	450	500	550	600	650	700
Y	800	900	1000	1100	1200	1300	1400	1500	1600

8.6 Calculate the product moment coefficient of correlation between X and Y from the following data:

X	7	6	5	4	3	2	1
Y	18	10	14	12	10	6	8

What will be the coefficient of correlation between U and V if $U = 3X + 2$ and $V = 25 - 2Y$?

8.7 The age (X) and blood pressure (Y) of nine women are given below. Find out correlation coefficient between X and Y:

X	57	59	62	63	64	65	65	58	57
Y	113	117	126	126	130	129	111	116	112

8.8 The deviations from their means of two series X and Y are respectively as given below.Find out the correlation coefficient:

X	-4	-3	-2	-1	0	1	2	3	4
Y	3	-3	-4	0	4	1	2	-2	-1

8.9 Calculate Karl Pearson's coefficient of correlation from the following data, using 20 as the working mean (assumed mean) for price and 70 as the working (assumed mean) for demand:

Price	14	16	17	18	19	20	21	22	23
Demand	84	78	70	75	66	67	62	58	60

8.10 Calculate Karl Pearson's coefficient of correlation between marks secured by 12 students in two tests:

Student	1	2	3	4	5	6	7	8	9	10	11	12
Test I	50	54	56	59	60	62	61	65	67	71	71	74
Test II	22	25	34	28	26	30	32	30	28	34	36	40

8.11 Calculate correlation coefficient between age and death-rate from the data given below:

Age Group	0-20	20-40	40-60	60-80	80-100
Death-rate	350	280	540	760	900

8.12 Calculate Karl Pearson's correlation coefficient between age and literacy from the data given below:

Age	Total Population ('000)	Literacy Population ('000)
10-20	120	100
20-30	100	75
30-40	80	60
40-50	60	40
50-60	40	25
60-70	15	10
70-80	8	4

8.13 Calculate correlation coefficient between age and success in examination:

Age	Candidates Approved	Successful Candidates
13-14	200	124
14-15	300	180
15-16	100	65
16-17	50	34
17-18	150	99
18-19	400	252
19-20	250	145
20-21	150	81
21-22	25	12
22-23	75	33

8.14 Find Karl Pearson's correlation coefficient between age and playing habits of the students from the following data:

Age (in years)	No. of Students	Regular players
15	250	200
16	200	150
17	150	90
18	120	48
19	100	30
20	80	12

8.15 Calculate the percentage of illiterate population from the following data and find out the correlation between age and illiteracy:

Age Group	Total Population ('000)	Illiterate Population ('00)
10-20	120	100
20-30	100	75
30-40	80	60
40-50	50	30
50-60	25	20
60-70	15	10
70-80	5	5

8.16 Calculate coefficient of correlation between density of population and death-rate from the following data:

Region	Area (sq.km)	Population	Deaths
1	200	40000	480
2	150	75000	1200
3	120	72000	1080
4	80	20000	280

8.17 The following table gives the distribution of students and also regular players among them according to age groups. Is there any correlation between age and playing habit?

Age	15-16	16-17	17-18	18-19	19-20	20-21
No. of students	200	270	340	360	400	300
Regular Players	150	162	170	180	180	120

8.18 From the following data calculate the coefficient of correlation by Karl Pearson's method:

X	6	2	10	?	8
Y	9	11	?	8	7

Arithmetic mean of X and Y-series are 6 and 8 respectively.

8.19 Calculate Karl Pearson's correlation coefficient from the following data by taking deviations from actual mean 8 (X-series) and 11 (Y-series):

X	11	10	9	?	7	6	5
Y	20	18	?	8	10	5	4

8.20 A few items are missing the following two series. Calculate Karl Pearson's coefficient of correlation. The additional information given is sum of the products of the deviations from the their respective means , $\sum xy = 38.4$, standard deviation of $A = 2$ and standard deviation of $B = 3$:

A	15	18	-	22	-	-	16	25
B	22	25	20	-	14	12	-	-

8.21 A few items are missing in the following two series but the additional information given is $\sum xy = -16, \sigma_x = 2.3, \sigma_y = 2.7$, where x = deviations from the $\overline{X}$ and y = deviations from $\overline{Y}$.

X	48	50	-	49	51	-	53	49
Y	36	-	33	38	-	-	55	30

8.22 In a set of 50 paired observations on two variables x and y , the following data were obtained: $\bar{x} = 10, \sigma_x = 3, \bar{y} = 6, \sigma_y = 2, r_{xy} = 0.3$.But on subsequent verification, it was found that one pairs of values $[x(=10), y(=6)]$ was inaccurate and hence rejected . With the remaining 49 pairs of observations,how the value of r_{xy} is affected?

8.5 INTERPRETATION OF CORRELATION COEFFICIENT

The interpretation of correlation coefficient depends on its degree and significance. The correlation coefficient lies between -1 and $+1$. For details go through the heading Degrees of correlation. The value of correlation coefficient is significant or not is judged with the help of probable error or standard error.

8.5.1 Computation of Correlation Coefficient in a Bivariate Frequency Distribution

In a bivariate frequency table (or a correlation table) one variable is written in column and the other is written in row. With no loss of generality we assume that the distribution of X is give in column and the distribution of Y is given in row. The following procedure is applied while calculating correlation coefficient when various groups are used.

Steps:.

1. Make three columns in L.H.S, three column in R.H.S , three rows in the top and three rows in the bottom of the given correlation table (excluding the column or row for classes).
2. Find the mid-values X, deviations from assumed or step-deviations d_x as the case may be and put them in the first three columns respectively.
3. Find the mid-values of Y, deviations from assumed mean or step-deviations d_y as the case may be and put them in first three rows respectively.
4. Multiply $d_x$by respectively frequency and get $\sum fd_x$.
5. Multiply $fd_x$by respective $d_x$and get $\sum fd_x^2$.
6. Multiply $d_y$by respective frequency and get $\sum fd_y$.
7. Multiply $fd_y$by respective d_y and get $\sum fd_y^2$.
8. Multiply respective $d_x$and $d_y$for each cell frequency and write the figures in the left hand upper corner of each cell.
9. To find fd_xd_y,multiply each cell frequency by corresponding $d_xd_y$and write these products in the right hand lower corner of each cell and get $\sum fd_xd_y$.
10. Finally use the following formula:

$$r = \frac{N\sum fd_xd_y - \sum fd_x \times \sum fd_y}{\sqrt{[N\sum fd_x^2 - (\sum fd_x)^2][N\sum fd_y^2 - (\sum fd_y)^2]}}$$

■ **EXAMPLE 8.11**

Calculate the coefficient of correlation between ages of husbands and ages of wives in the following bivariate frequency distribution. Find also its probable error and comment on the result:

Ages of husbands	Ages of Wives				
	10-20	20-30	30-40	40-50	50-60
15-25	6	3	_	_	_
25-35	3	16	10	_	_
35-45	_	10	15	7	_
45-55	_	_	7	10	4
55-65	_	_	_	4	5

Solution:

Ages of husbands (Y)	$d_y \mid d_x$	Ages of Wives (X) 10-20 -2	20-30 -1	30-40 0	40-50 1	50-60 2	f	fd_y	fd_y^2	fd_xd_y
15-25	-2	6(24)	3(6)	_	_	_	9	-18	36	30
25-35	-1	3(6)	16(16)	10(0)	_	_	29	-29	29	22
35-45	0	_	10(0)	15(0)	7(0)	_	32	0	0	0
45-55	1	_	_	7(0)	10(10)	4(8)	21	21	21	18
55-65	2	_	_	_	4(8)	5(20)	9	18	36	28
	f	9	29	32	21	9	N=100			
	fd_x	-18	-29	0	21	18				
	fd_x^2	36	29	0	21	36				
	fd_xd_y	30	22	0	18	28				

$$d_x = \frac{x-35}{10},\ d_y = \frac{y-40}{10}$$

[Table Interpretation: We first obtain mid values of classes as X and Y for both variables. Then, we compute d_x and d_y. The values given in brackets show the product of d_x , d_y and f representing that cell, i.e., fd_xd_y. For example for first cell $d_x = -2$ and $d_x = -2$, therefore $d_xd_xy = 4$ and frequency of that cell is 6 therefore $fd_xd_y = 24$, which is written in bracket there. Finally, look the last column for the first row; there fd_xd_y is sum of all fd_xd_y in that row. similarly in other row and columns.]

Here, $\sum fd_x = -8$, $\sum fd_y = -8$, $\sum fd_x^2 = 122$, $\sum fd_y^2 = 122$, and $\sum fd_xd_y = 98$. Hence

$$r = \frac{N\sum fd_xd_y - \sum fd_x \times \sum fd_y}{\sqrt{[N\sum fd_x^2 - (\sum fd_x)^2][N\sum fd_y^2 - (\sum fd_y)^2]}}$$

$$r = \frac{100\times 98 - (-8)\times(-8)}{\sqrt{[100\times 122 - (-8)^2][100\times 122 - (-8)^2]}}$$
$$= 0.802$$

High degree positive correlation.
Probable Error,

$$\text{P.E.} = 0.6745 \times \frac{1-r^2}{\sqrt{N}} = 0.6745 \times \frac{1-(0.802)^2}{\sqrt{100}} = \frac{0.6745*(1-0.64)}{10} = 0.06745\times 0.36 = 0.024$$

$$\therefore 6(P.E.) = 6\times 0.024 = 0.144$$

Since $r > P.E.$, correlation is significant. □

Problems

8.23 Calculate coefficient of correlation from the following data:

x y →	-3	-2	-1	0	1	2	3
-3	–	–	–	–	–	–	10
-2	–	–	–	–	16	6	8
-1	–	–	–	10	14	8	–
0	–	–	4	10	18	6	–
1	–	4	6	12	–	–	–
2	6	6	–	–	–	–	–
3	8	–	–	–	–	–	–

8.24 The following figures are obtained in connection with the income and saving survey of 100 school principals in a city:

Income (in Rs.)	50	120	150	200	Total
400	8	4	–	–	12
600	–	12	24	6	42
800	–	9	7	2	18
1000	–	–	10	5	15
1200	–	–	9	4	13
Total	8	25	50	17	100

Find correlation between income and saving.

8.25 Calculate Karl Pearson's correlation coefficient from the following table:

x\y	18	19	20	21	22	Total
0-5	6	1	–	–	–	7
5-10	–	12	3	–	5	20
10-15	–	3	–	–	2	5
15-20	–	–	–	4	2	6
20-25	–	4	2	2	3	11
Total	6	20	5	6	12	49

8.26 Determine Karl Pearson's correlation coefficient from the following bivariate frequency distribution:

x\y	20-25	25-30	30-35	35-40	40-45
25-30	3	–	–	–	–
30-35	–	2	2	–	–
35-40	–	5	3	3	–
40-45	–	–	3	4	–
45-50	–	–	1	–	4

8.6 RANK CORRELATION

When the variables under consideration are not capable of quantitative measurement but can be arranged in serial order (ranks), we find correlation between the ranks of two series. This happens when we deal with qualitative characteristics such as honesty, beauty, etc. This method is called *Spearman's Rank Difference Method* or Ranking Method and the correlation coefficient so obtained is called Rank Correlation Coefficient and is denoted by r_s. This method was developed by Charles Edward Spearman, a British Psychologist in 1904.

Spearman's rank correlation coefficient is also used when the measurements are given for both the series.

However, Spearman's rank correlation coefficient r_s is nothing but Karl Pearson's correlation coefficient between the ranks , it can be interpreted in the same way as the Karl Pearson's correlation coefficient.

Case 1. When ranks are different in the two series , then

$$r_s = 1 - \frac{6\sum D^2}{N(N^2-1)} \tag{8.1}$$

where, r_s =coefficient of rank correlation

$\sum D^2$= Sum of squares of the differences in ranks ($D = R_x - R_y$, R_x, R_ydenotes rank in x and y data.)

N = Number of Pairs

Case 2:

When there is a tie i.e., if any two or more individuals have equal ranks then the formula (1) for calculating rank correlation coefficient breaks down we use the following formula:

$$r = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m^3-m) + ...]}{N(N^2-1)} \tag{8.2}$$

Where m is the number of times an item is repeated. The correction factor $\frac{1}{12}(m^3-m)$ is to be added for each repeated value in both the series.

8.6.1 Computation of Rank correlation Coefficient

Steps (when ranks are given).

1. Compute $D = R_x - R_y$, the difference of ranks in the two series.
2. Compute $D^2$and get $\sum D^2$.
3. Use the formula (8.1) or (8.2) as the case may be .

Steps (when ranks are not given).

1. Convert the given values into ranks separately for both series;
   (a) The highest observation is given the rank 1. The next highest observation is given rank 2 and so on. (However the rank 1 may be given to smallest observation,etc.)

(b) In case of a tie, we may use 'bracket rank method' or 'average rank method'. In bracket rank method , we give equal ranks to the like items but the next item is given the rank as it would be in case of no ties. For example ,

Consider the series $20, 25, 22, 21, 22$. Here 25 is given rank 1, both 22 are given rank 2 and 2, 21 is given rank 4 (Note: Now rank 3 is not there.), 20 is given rank 5.

In average rank method, common rank to be assigned to each item is the average of the ranks which these observations would have assumed if they were different.

Consider the series $25, 25, 30, 24, 25, 27, 35$. Here 35 is given rank 1, 30 is given rank 2, 27 is given rank 3, 25 is given rank $\frac{4+5+6}{3}$ i.e., 5 for each, 24 is given rank 7 (Note: Now rank 4 and 6 are not there.).

2. Compute D, the difference of ranks.
3. Compute $D^2$and get $\sum D^2$.
4. Use the formula (8.1) or (8.2) as the case may be .

EXAMPLE 8.12

Ten competitors in a beauty contest are ranked by three judges in the following order:

First Judge	1	6	5	10	3	2	4	9	7	8
Second Judge	3	5	8	4	7	10	2	1	6	9
Third Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to discuss which pair of judges have the nearest approach to common tastes in beauty ?

Solution: Here we calculate three rank correlation coefficient:
r_{12} between the ranks of judges I and II
r_{23} between the ranks of judges II and III
r_{13} between the ranks of judges I and III.

Judge I	Judge II	Judge III	D_{12}	D_{12}^2	D_{23}	D_{23}^2	D_{13}	D_{13}^2
R_1	R_2	R_3	$R_1 - R_2$		$R_2 - R_3$		$R_1 - R_3$	
1	3	6	-2	4	-3	9	-5	25
6	5	4	1	1	1	1	2	4
5	8	9	-3	9	-1	1	-4	16
10	4	8	6	36	-4	16	2	4
3	7	1	-4	16	6	36	2	4
2	10	2	-8	64	8	64	0	0
4	2	3	2	4	-1	1	1	1
9	1	10	8	64	-9	81	-1	1
7	6	5	1	1	1	1	2	4
8	9	7	-1	1	2	4	1	1
Total	-	-	0	200	0	214	0	60

Formula:

$$r_{12} = 1 - \frac{6\sum D_{12}^2}{N(N^2-1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - 1.212 = -0.212$$

$$r_{23} = 1 - \frac{6\sum D_{23}^2}{N(N^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = 1 - 1.297 = -0.297$$

$$r_{13} = 1 - \frac{6\sum D_{13}^2}{N(N^2-1)} = 1 - \frac{6 \times 60}{10 \times 99} = 1 - 0.364 = +0.636$$

The Judges I and III have nearest approach to common tastes in beauty where as the Judges II and III disagree the most. □

■ EXAMPLE 8.13

Find out the coefficient of correlation between X and Y by the method of rank differences:

X	20	22	24	25	30	32	28	21	26	35
Y	16	15	20	21	19	18	22	24	23	25

Solution:

X	R_x	Y	R_y	$D = R_x - R_y$	D^2
20	10	16	9	+1	1
22	8	15	10	-2	4
24	7	20	6	+1	1
25	6	21	5	+1	1
30	3	19	7	-4	16
32	2	18	8	-6	36
28	4	22	4	0	0
21	9	24	2	+7	49
26	5	23	3	+2	4
35	1	25	1	0	0
$N = 10$		$N = 10$			$\sum D^2 = 112$

$$r_s = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6 \times 112}{10(10^2-)} = 1 - \frac{672}{10(100-1)}$$

$$r_s = 1 - \frac{672}{10 \times 99} = 1 - \frac{672}{990} = \frac{990-672}{990} = \frac{318}{990} = 0.32$$

Positive Correlation of low degree. □

■ EXAMPLE 8.14

Find out the coefficient of correlation between X and Y by the method of rank differences:

X	22	24	27	35	21	20	27	25	27	33
Y	30	28	40	50	38	25	38	36	41	32

Solution:

X-Series	Rank R_x	Y-Series	Rank R_y	Difference of Rank (D)	D^2
22	8	30	9	-1	1
24	6	38	5	+1	1
27	3	40	3	0	0
35	1	50	1	0	0
21	9	38	5	+4	16
20	10	25	10	0	0
27	3	38	5	-2	4
25	5	36	7	-2	4
27	3	41	2	+1	1
23	7	32	8	-1	1
$N = 10$		$N = 10$			$\sum D^2 = 28$

Formula:

$$r_s = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)]}{N(N^2-1)}$$
$$= 1 - \frac{6[28 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(3^3 - 3)]}{10(10^2-1)}$$
$$= 1 - \frac{6[28 + \frac{1}{12}(27 - 3) + \frac{1}{12}(27 - 3)]}{10(100-1)}$$
$$= 1 - \frac{6(28+2+2)}{10(99)} = +0.81$$

□

■ EXAMPLE 8.15

Find out the value of r_s:

X	15	14	25	14	14	20	22
Y	25	12	18	25	40	10	7

Solution:

X	R_x	Y	R_y	$D = R_x - R_y$	D^2
15	4	25	2.5	1.5	2.25
14	6	12	5	1	1
25	1	18	4	-3	9
14	6	25	2.5	3.5	12.25
14	6	40	1	5	25
20	3	10	6	-3	9
22	2	7	7	-5	25
$N = 7$					$\sum D^2 = 83.50$

Formula:

$$
\begin{aligned}
r_s &= 1 - \frac{6[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)]}{N(N^2-1)} \\
&= 1 - \frac{6[83.5 + \frac{1}{12}(3^3-3) + \frac{1}{12}(2^3-2)]}{7(7^2-1)} \\
&= 1 - \frac{6(83.5+2+.5)}{7(49-1)} = 1 - \frac{6 \times 86}{7(48)} = -0.54
\end{aligned}
$$

□

■ EXAMPLE 8.16

The coefficient of rank correlation of the marks obtained by 10 students in Mathematics and Statics was found to be 0.91. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 0 instead of 3. Find the correct coefficient of rank correlation.

Solution:

$$
\begin{aligned}
r_s &= 1 - \frac{6\sum D^2}{N(N^2-1)} \\
\Rightarrow 0.91 &= 1 - \frac{6\sum D^2}{10(10^2-1)} \\
\Rightarrow 0.91 &= 1 - \frac{6\sum D^2}{990} \\
\Rightarrow 0.91 \times 990 &= 990 - 6\sum D^2 \\
\Rightarrow \sum D^2 &= 15
\end{aligned}
$$

$$\text{Correct } \sum D^2 = 15 + (3)^2 - 0^2 = 24$$

$$\therefore \text{Correct, } r_s = 1 - \tfrac{144}{990} = \tfrac{846}{990} = .85$$

□

■ EXAMPLE 8.17

From the following data calculate Spearman's Rank coefficient of correlation:

Serial Number	1	2	3	4	5	6	7	8	9	10
Rank Difference	-2	-4	-1	+3	+2	0	?	+3	+2	-2

Solution: Let the unknown rank be a. Since the sum of the difference is zero i.e., $\sum D = 0$,we have

$$-2-4-1+3+2+0+a+3+3-2=0$$

$$-9+11+a=0$$

$$\therefore a = -2$$

Hence

$$\sum D^2 = 4+16+1+9+4+0+4+9+9+4 = 60$$

and

$$r_s = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6\times 60}{10\times 99} = 1 - 0.36 = 0.64$$

□

8.6.2 Characteristics of Spearman's Correlation Method

1. Spearman's formula is simple to calculate and easy to understand.
2. It is the only formula to be used for finding correlation coefficient if we are dealing with qualitative data.
3. It can also be used in case of extreme values where observations are irregular.
4. It is not practicable in case of bivariate frequency distribution.
5. When number of observations is more than 30,then this formula is not suitable.
6. In Spearman's formula: $\sum d = 0$ always, $\sum d^2$ =even (in case of untied ranks),$-1 < r_s < +1$, minimum value of $\sum d^2 = \frac{N^3-N}{3}$.

Problems

8.27 Following are the ranks obtained by 10 students in two , Statistics and Mathematics.To what extent the knowledge of the students in the two subjects in the subjects is related ?

Rank in Statistics	1	2	3	4	5	6	7	8	9	10
Rank in Mathematics	2	4	1	5	3	9	7	10	6	8

8.28 Ten girl students were ranked according to their performance in a voice test by two Judges as given below:

Girls	1	2	3	4	5	6	7	8	9	10
Judge I	4	8	6	7	1	3	2	5	10	9
Judge II	3	9	6	5	1	2	4	7	1	10

Are the likings of two Judges with regard to the voice similar ?

8.29 The ranks of the same 15 students in two subjects A nd B are given below ; the two numbers within the brackets denoting the ranks of the same student in A and B respectively:

$(1,10),(2,7),(3,2),(4,6),(5,4),(6,8),(7,3),(8,1),(9,11),(10,15),(11,9),(12,5),(13,14),(14,12),(15,13)$

Use Spearman's formula to find the rank correlation coefficient.

8.30 Calculate Spearman's rank correlation coefficient between ranks (R_x)of the profit earned (X) and the ranks (R_y)of working capital (Y) for eight industries:

R_x	1	8	7	6	5	4	3	2
R_y	8	1	3	4	2	6	5	7

8.31 Calculate the coefficient of correlation from the following data by Spearman's rank differences method:

Price of Tea (Rs.)	75	88	95	70	60	80	81	50
Price of Coffee (Rs.)	120	130	150	115	110	140	142	100

8.32 Calculate rank correlation coefficient from the following data:

X	20	25	30	35	40	45	50	55	60	65	70
Y	17	24	28	32	35	30	29	51	56	60	62

8.33 Calculate Spearman's rank correlation coefficient between advertisement cost and sales from the following data:

Cost ('000 Rs.)	48	33	40	9	16	16	65	24	16	57
Sales (lakh Rs.)	13	13	24	6	15	4	20	9	6	19

8.34 From the following data , find the coefficient of correlation by the method of rank differences:

X	52	63	45	36	72	65	45	25
Y	62	53	51	25	79	43	60	32

8.35 Calculate correlation coefficient from the following data by rank difference method:

Students	A	B	C	D	E	F	G	H	I	J
Marks in English	22	27	28	15	32	14	7	9	5	6
Maths	32	39	40	20	21	33	16	35	48	10

8.36 Calculate correlation coefficient from the following data by Spearman's Ranking Method ;

X	80	78	75	75	68	67	60	59
Y	110	111	114	114	114	116	115	117

8.37 Calculate correlation coefficient by the method of rank differences from the data given below:

X	115	109	112	87	98	98	120	100	98	118
Y	75	73	85	70	76	65	82	73	68	80

8.38 Use the method of rank correlation to determine the relationship between Preference prices (X) and Debenture price (Y) from the data given below:

X	73.2	85.8	78.9	75.8	77.2	87.2	83.8
Y	97.8	99.2	98.8	98.3	98.4	96.7	97.1

CHAPTER 9

REGRESSION

9.1 REGRESSION ANALYSIS

The Dictionary meaning of regression is 'Stepping back'. The term regression was first used by a British Biometrician Sir Francis Galton (1822-1911) in the latter past (1877) of the nineteenth century in connection with the height of parents and their offsprings. He found that the offsprings of tall or short parents tend to regress to the average height. In other words, the average height of tall fathers is greater than the average height of their offsprings and the average height of short fathers is less than the average height of their offsprings. Galton saw in this a tendency of the human race to regress or return a normal height. He referred to this tendency to return to the average height of all parents as regression in his research paper "Regression toward mediocrity in hereditary stature". But now-a-days the term 'regression' stands for some sort of relationships between two or more variables. The variable which is used to predict the variable of interest is called the independent variable or explanatory variable and the variable which is predicted or estimated is called the dependent variable or explained variable.

regression analysis refers to techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps us understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

9.1.1 Utility of Regression

In brief, the utility of regression or regression analysis can be understand in the following heads:

1. Estimation of Values: When there is a cause and effect relationship between two variables. The one can be estimated with the help of the other.

2. Determination of the rate of change in the dependent variable for a unit change in the independent variable.

3. Measurement of the errors involved

4. Measurement of degree and direction of correlation: We can measure the degree and direction of correlation with the help of regression analysis.

9.1.2 Difference Between Correlation and Regression

The points of difference between correlation and regression are as follows:

1. While correlation analysis tests the closeness with which two (or more) phenomena covary, regression analysis measures the nature and extent of the relation, thus enabling us to make predictions.

2. Correlation analysis has limited applications as it is confined only to the study of relationship between variables; but the regression analysis has much wider application because it also predicts.

3. In the cause-effect relationship, correlation analysis does not consider the concept of dependent and independent variables. On the contrary, in regression analysis one variable is assumed as dependent which the others an independent.

4. In correlation analysis, the correlation coefficient is a measure of degree of co-variability between two variables while the regression analysis deals with a functional relationship between two variables.

Conclusion. A good way of distinguishing between correlation and regression is to consider the correlation phase of study of variable to determine the strength of their relationship and to view regression as that phase which deals with prediction of one variable on the basis of its correlation with other. Naturally correlation should precede regression for if the relationship is not sufficiently strong, there would appear to be no sound basis for prediction.

9.1.3 Kinds of Regression Analysis

The regression analysis may be classified as:
(i) Linear and Curvilinear Regression,
(ii) Simple and Multiple Regression.

Linear Regression . If the relationship can be represented by a straight line, then it is known as linear regression. In this case the values of the dependent variable changes at a constant rate for a unit change in the values of the independent variable.

Non-liner Regression . When the relationship between variables can be represented by a curve other than straight line, then it is called Non-linear regression.

Simple Regression . If the study is based only on two variables, it is called simple regression.

Multiple Regression . If the study is based on more than two variables, it is called multiple regression.

9.1.4 Regression Lines

The lines of best fit expressing mutual average relationship between two variables are known as regression lines. For two variables, there are two regression lines. Let us consider two variables X and Y. If we have no reason or justification to assume as dependent variable and other as independent variable either of the two may be taken as independent.

1. If we wish to estimate Y for given values of X, we take X as independent variable and Y as dependent variable. In this case the straight line is called as the regression line of Y on X.
2. If we wish to estimate X for given values of Y, we take Y as independent variable and X as dependent variable. In this case the straight line is called as the regression line of X on Y.

The concept of lines of best fit is based on the principle of least squares. In a scatter diagram we take X on x-axis. When the deviations from the points to the line of best fit are measured horizontally i.e., parallel to x-axis and the sum of squares of these deviations is minimized, we get, the regression line of X on Y. When the deviations from the points to the line of best fit are measured vertically i.e., parallel to y-axis and the sum of squares of these deviations is minimized, we get, the regression line of Y on X. Thus, in general, we have always two lines of regression in case of simple regression (i.e., for variables). The two regression lines cut each other at the point $(\bar{x}, \bar{y})$.

9.1.4.1 Regression Lines and Correlation

1. If there is either perfect positive of perfect negative correlation between the two variables (i.e., $r = \pm 1$), the two regression lines coincide and there will be only one line of regression.
2. If the two regression lines cut each other at right angle, then $r = 0$.
3. The nearer the two regression lines (i.e., the smaller the acute angle between the two regression lines) the greater is the degree of correlation. The various situations are shown in the figure 1.

9.1.4.2 Construction of Regression Lines There are two methods of drawing regression lines: (i) the two regression lines can be drawn by free hand method, (ii) with the help of regression equations.

Regression Coefficients. Let r be the correlation Coefficient between X and Y, σ_x be standard deviation of X and σ_y be standard deviation of Y. Then

(a) $r\frac{\sigma_x}{\sigma_y}$ is called the regression coefficient of X on Y and is denoted by b_{xy}.

(b) $r\frac{\sigma_y}{\sigma_x}$ is called the regression coefficient of Y on X and is denoted by b_{yx}.

Thus

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \text{ and } b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$\Rightarrow \sqrt{b_{yx} \times b_{xy}} = \sqrt{r\frac{\sigma_y}{\sigma_x}.r\frac{\sigma_x}{\sigma_y}} = \pm r$$

The sign of b_{yx} and b_{xy} are same. When both are positive, then r is positive, when both are negative then r is negative.

Regression of Y on X. The line of regression of Y on X is

$$(Y - \overline{Y}) = b_{YX}(X - \overline{X})$$

where $\overline{X}, \overline{Y}$ represents the mean of X and Y respectively. b_{YX} is regression coefficient of Y on X and is given as

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

To compute b_{YX}, we use following formula:

$$b_{YX} = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$$

or

$$b_{YX} = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2}$$

where $x = X - A$ and $y = Y - B$, A and B are assumed mean. or

$$b_{YX} = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2} \times \frac{k}{h}$$

where $x = \frac{X-A}{h}$ and $y = \frac{Y-B}{k}$, A and B are assumed mean.

Regression of X on Y. The line of regression of X on Y is

$$(X - \overline{X}) = b_{XY}(Y - \overline{Y})$$

where $\overline{X}$, $\overline{Y}$ represents the mean of X and Y respectively. b_{XY} is regression coefficient of X on Y and is given as

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

To compute b_{YX}, we use following formula:

$$b_{XY} = \frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - (\sum Y)^2}$$

or

$$b_{XY} = \frac{N\sum xy - \sum x \sum y}{N\sum y^2 - (\sum y)^2}$$

where $x = X - A$ and $y = Y - B$, A and B are assumed mean. or

$$b_{YX} = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2} \times \frac{h}{k}$$

where $x = \frac{X-A}{h}$ and $y = \frac{Y-B}{k}$, A and B are assumed mean.

■ EXAMPLE 9.1

Find the most likely price at Delhi corresponding to the price of Rs. 70 at Agra from the following data:

	Agra	Delhi
Average Price (Rs.)	65	67
Standard Deviation	2.5	3.5

Coefficient of correlation between the prices of the two places +0.8.

Solution:

Let x = price at Agra, y = price at Delhi
Given $x = 65, \overline{y} = 67, \sigma_x = 2.5, \sigma_y = 3.5, r = +0.8$
The regression equation of y on x

$$y - \overline{y} = r\frac{\sigma_y}{\sigma_x}(x - \overline{x})$$

Substituting the known values,

$$y - 67 = 0.8 \times \frac{3.5}{2.5}(70 - 65) = 5.6$$

$$\therefore y = 5.6 + 67 = 72.6$$

$\therefore$ The most likely price at Delhi is 72.6 when the price at Agra is 70.

□

■ EXAMPLE 9.2

An investigation into the demand of Television Sets in 7 Towns has resulted in the following data:

Population (in thousand)	x:	11	14	14	17	17	21	25
No. of T.V. Sets demanded	y:	15	27	27	30	34	38	46

Calculate regression equation of y on x, and estimate the demand for T.V. Sets for a town with a population of 30 thousand.

Solution:

Population X	$x = (X - 17)$	x^2	T.V. Sets Y	$y = (Y - 31)$	y^2	xy
11	-6	36	15	-16	256	+96
14	-3	9	27	-4	16	+12
14	-3	9	27	-4	16	+12
17	0	0	30	-1	1	0
17	0	0	34	+3	9	0
21	+4	16	38	+7	49	28
25	+8	64	46	+15	225	120
$\Sigma X = 119$	$\sum x = 0$	$\Sigma x^2 = 134$	$\Sigma Y = 217$	$\sum y = 0$	$\Sigma y^2 = 572$	$\Sigma xy = 268$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{119}{7} = 17 \text{ and } \bar{Y} = \frac{\Sigma Y}{N} = \frac{217}{N} = 31$$

Now prepare table. Then

$$b_{YX} = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2}$$
$$= \frac{7 \times 3957 - 119 \times 217}{7 \times -(119)^2} = 2$$

Therefore, the Regression equation of Y on X:

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

Substituting the known values,

$$Y - 31 = 2(X - 17) \ \Rightarrow \ Y = 2X - 3$$

To obtain the average demand of T.V. Sets when the population is 30 thousand, we substitute 30 in place of X in the regression equation of Y on X, then we have:

$$y = (2 \times 30) - 3 = 57$$

□

9.1.5 Regression Equations using Method of Least Squares

Let the regression equation of x on y be $x = c + dy$, then normal equations are

$$\Sigma x = nc + d\Sigma y$$

$$\Sigma xy = c\Sigma y + d\Sigma y^2$$

Find the value of $\Sigma y, \Sigma x, \Sigma xy$and Σy^2; substitute these values in normal equations and solve them for c and d. Let the regression equation of y on x be $y = a + bx$, then the normal equations are

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Find the values of $\Sigma x, \Sigma y, \Sigma xy$and Σx^2; substitute these values in normal equations and solve them. **Remark:** For simplicity one may take deviations from assumed mean followed by necessary adjustments.

EXAMPLE 9.3

From the following data, obtain the two regression equations:

X	6	2	10	4	8
Y	9	11	5	8	7

Solution:

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
$\sum X = 30$	$\sum Y = 40$	$\sum X^2 = 220$	$\sum Y^2 = 340$	$\sum XY = 214$

Regression of Y on X. Let the line of regression of Y on X

$$Y = a + bX$$

Then normal equations are

$$\sum Y = Na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

putting the values from table, we get

$$40 = 5a + 30b$$

$$214 = 30a + 220b$$

On solving these equations, we get

$$a = 11.9, \text{ and} b = -0.65$$

Hence line of regression is

$$Y = 11.9 - 0.65X$$

Regression of Y on X. Let the line of regression of Y on X

$$X = m + nX$$

Then normal equations are

$$\sum X = Nm + b\sum Y$$
$$\sum XY = m\sum Y + b\sum Y^2$$

putting the values from table, we get

$$30 = 5m + 40n$$
$$214 = 40m + 340n$$

On solving these equations, we get

$$m = 16.4, \text{ and} n = -1.3$$

Hence line of regression is

$$X = 16.4 - 1.3Y$$

Remark: Students may also solve this problems by procedure of previous example. □

Regression Lines for Grouped Data. For grouped data regression coefficients are given as below:
Regression coefficient of X on Y

$$b_{XY} = \frac{N\sum fxy - \sum fx\sum fy}{N\sum fy^2 - (\sum fy)^2} \times \frac{h}{k}$$

Regression coefficient of Y on X

$$b_{YX} = \frac{N\sum fxy - \sum fx\sum fy}{N\sum fx^2 - (\sum fx)^2} \times \frac{k}{h}$$

Other procedure remains same. Read following example carefully.

■ EXAMPLE 9.4

Following table gives the ages of husbands and wives for 50 newly married couples. Find the two regression lines. Also estimate (a) the age of husband when wife is 20, and (b) the age of wife when husband is 30.

Age of Wives	Age of Husbands			Total
	20-25	25-30	30-35	
16-20	9	14	-	23
20-24	6	11	3	20
24-28	-	-	7	7
Total	15	25	10	50

Width of the class-interval for age of husband (x) is $h = 5$
Width of the class interval for age of wife (y) is $k = 4$

Y	Y_{Mid}	X / X_{Mid} / $y\backslash x$	20-25 / 22.5 / -1	25-30 / 27.5 / 0	30-35 / 32.5 / +1	Total	fy	fy^2	fxy
16-20	18	-1	9 (9)	14(0)		23	-23	23	9
20-24	22	0	6 (0)	11 (0)	3 (0)	20	0	0	0
24-28	26	+1			7(7)	7	7	7	7
		Total	15	25	10	50	-16	30	16
		fx	-15	0	10	-5			
		fx^2	15	0	10	25			
		fxy	9	0	7	16			

Regression coefficient of X on Y

$$b_{XY} = \frac{N\sum fxy - \sum fx \sum fy}{N\sum fy^2 - (\sum fy)^2} \times \frac{h}{k}$$

or

$$b_{xy} = \frac{16 \times 50 - (-5)(-16)}{30 \times 50 - (-16)^2} \times \frac{5}{4} \approx 0.723$$

Regression coefficient of Y on X

$$b_{YX} = \frac{N\sum fxy - \sum fx \sum fy}{N\sum fx^2 - (\sum fx)^2} \times \frac{k}{h}$$

or

$$b_{YX} = \frac{16 \times 50 - (-5)(-16)}{25 \times 50 - (-5)^2} \times \frac{4}{5} \approx 0.47$$

$$\bar{X} = A + \frac{\Sigma fx}{N} \times h$$

$$= 27.5 + \frac{-5}{50} \times 5 = 27$$

and

$$\bar{Y} = A + \frac{\Sigma fy}{N} \times k$$

$$= 22 + \frac{-16}{50} \times 4 = 20.72$$

Regression equation of X on Y:

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$X - 27 = 0.723(Y - 20.72) \ \Rightarrow \ X = 0.723Y + 12.02$$

(a) Estimate of husband's age
Substitute $Y = 20$ in

$$X = 0.723y + 12.02$$

$$\Rightarrow x = 0.723 \times 20 + 12.02 = 26.48 \text{ Years}$$

Regression equation of Y on X:

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$Y - 20.72 = 0.47(X - 27) \ \Rightarrow \ Y = 0.47X + 8.03$$

(b) Estimate of wife's age
Substitute $X = 30$ in

$$Y = .47X + 8.03$$

$$\Rightarrow Y = 0.47 \times 30 + 8.03 = 22.13 \text{ Years}$$

9.1.6 Some Facts about Regression Equations and Regression Coefficient

1. The regression lines of y on x and x on y both pass through the points $(\bar{x},\bar{y})$.That is the two regression lines intersect each other at the point $(\bar{x},\bar{y})$.
2. If correlation coefficient r between x and y is zero (r=0), the two regression lines are $y=\bar{y}$ and $x=\bar{x}$ respectively. (i.e., the two regression lines are perpendicular to each other)
3. If $r=+1$or $r=-1$, then the two equations reduce to
$$\frac{x-\bar{x}}{\sigma_x}=\pm\frac{y-\bar{y}}{\sigma_y}$$
i.e., the two equations of regression coincide and there is only one regression equation.
4. The geometric mean of the two regression coefficients $b_{yx}$and b_{xy} is equal to the correlation coefficient
$$\sqrt{b_{yx}\times b_{xy}}=\sqrt{r\frac{\sigma_y}{\sigma_x}.r\frac{\sigma_x}{\sigma_y}}=\sqrt{r^2}=r$$
The sign of the three $b_{yx},b_{xy}$and r will be the same.
5. The arithmetic mean of the regression coefficients is greater than correlation coefficient. i.e.,
$$\frac{b_{yx}+b_{xy}}{2}>\sqrt{b_{yx}b_{xy}}=\frac{b_{yx}+b_{xy}}{2}>r$$
6. The regression coefficient are independent of the change of origin but not of scale: Let $U=\frac{X-A}{H},V=\frac{Y-B}{K}$, then $b_{vu}=\frac{H}{K}b_{yx}$and $b_{uv}=\frac{K}{H}b_{xy}$
7. If one of the regression coefficient is greater than unity, the other must be less than unity provide $b_{yx}\times b_{xy}\leq 1$. (With the help of this property we decide which of the two given equations is y on x and which one is x on y.)
8. Regression coefficient is the algebraic measure of the slope of the regression line, it gives the change in the values of a series for a unit change in the values of the other series.

9.1.7 Miscellaneous examples based on Regression Coefficients

■ **EXAMPLE 9.5**

Find out the value of r when $b_{yx}=0.64$and $b_{xy}=1$.

$$r=\sqrt{b_{yx}\times b_{xy}}=\sqrt{0.64\times 1}=\sqrt{0.64}=0.8$$

■ **EXAMPLE 9.6**

If the two regression coefficients are -0.9 and -0.5. find out the value of correlation coefficient.

Since b_{yx} and b_{xy} are negative

$$r=-\sqrt{b_{xy}\times b_{xy},}$$

$$=\sqrt{(-0.9)(-0.5)}=-\sqrt{0.45}=-0.67$$

■ **EXAMPLE 9.7**

The two regression equations obtained by a student were as given below:

$$3x - 4y = 5,\ 8x + 16y = 15$$

Do you agree with him? Explain with reasons.

Let $3x - 4y = 5$ be regression line of x on y, then

$$x = \frac{4}{3}y + \frac{5}{3} \Rightarrow b_{xy} = \frac{4}{3}$$

Let $8x + 16y = 15$ be regression line of y on x, then

$$y = -\frac{8}{16}x + \frac{15}{16} \Rightarrow b_{yx} = -\frac{8}{16} = -\frac{1}{2}$$

Since b_{xy} is positive and b_{yx} is negative (i.e. the sign of the two regression coefficients is not same), we do not agree with him.

■ EXAMPLE 9.8

In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: Variance of $x = 9$, Regression equations are: (i) $8x - 10y = -66$; (ii) $40x - 18y = 214$. What are (a) the mean values of x and y, (b) the coefficient of correlation between x and y, (c) the σ of y.

(a) Means of x and y: Solve the equation (i) and (ii),

$$8x - 10y = -66 \tag{9.1}$$

$$40x - 18y = 214 \tag{9.2}$$

On solving equations 9.1 and 9.2,

$$x = 13,\ y = 17$$

(b) Correlation Coefficient:

$$8x - 10y + 66 = 0$$

$$40x - 18y - 214 = 0$$

Assume equation (i) as y on x,

$$-10y = -8x = -66$$

$$y = \frac{-8x - 66}{-10} = 0.8x + 6.6$$

$\therefore$Regression coefficient of y on x,

$$b_{yx} = 0.8$$

Assume equation (ii) as x on y,

$$40x = 18y + 214$$

$$x = \frac{18y + 214}{40} = 0.45y + 5.35$$

$\therefore$Regression coefficient of x on y,

$$b_{xy} = 0.45$$

$$\therefore r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{.45 \times .8} = \sqrt{.36} = +0.6$$

(c) Standard Deviation of y:

Here

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.8$$

$$0.8 = 0.6\frac{\sigma_y}{\sqrt{9}}$$

$$\sigma_y = \frac{2.4}{0.6} = 4$$

Problems

9.1 The following data related with the prices of items in Jabalpur and Narsinghpur are available. Estimate the price of a item in Narsinghpur when its price in Jabalpur is Rs.15:

	Jabalpur	Narsinghpur
Mean of the prices of the item	10	12
S.d. of the prices of the item	4.2	4.5

The correlation coefficient between the prices of item in Jabalpur and Narsinghpur = $+0.8$

9.2 The average weekly wages of working class in any places A and B are Rs. 12 and Rs. 18 respectively. Their standard deviations are Rs. 2 and Rs. 3 respectively and the coefficient of correlation between them is +0.67. Find out the most likely wage in place B if it is Rs. 20 in place A.

9.3 The following data based on 450 students are given for marks in Statistics and Economics at a certain examination. Estimate the average marks in Economics of the candidates who obtained 50 marks in Statistics:

	Statistics	Economics
Mean marks	40	48
Standard deviation	12	16

Sum of products of deviations of marks from their respectively means =42,075

9.4 Given the following data:

	Series A	Series B
Mean	60	75
Standard Deviation	4.4	2.2

r between A and $B = 0.8$.

(i) Obtain the two regression equations and find out the most probable value (a) of A when B is 50 and (b) of B when A is 80.

(ii) Draw the two regression lines on the graph and verify your results.

9.5 The following data are given for marks in subjects A and B in a certain examination:

	Subject A	Subject B
Mean of Marks	39.5	47.6
S.D. of Marks	10.8	16.9

Coefficient of correlation between A and $B = 0.42$.

Obtain the equation of the two regression line. Estimate the probable marks of subject A. if the marks of subject B are 55 and estimate the probable marks of subject B if the marks of subject A are 50.

9.6 Obtain the two lines of regression from the following data:

x	1	2	3	4	5
y	7	8	10	12	13

9.7 Following data refer to years service in a factory of seven persons in a specialized field and to their monthly income:

Years of Service	11	7	9	5	8	6	10
Income (in Rs. hundred)	7	5	3	2	6	4	8

Find the regression equation of income on years of service. Using it, what initial start would you recommend for a person applying for a job after having served in another factory in a similar field for 12 years?

9.8 Find the lines of regression for the two series given below. What is the most likely value of y when $x = 150$.

x	147	148	135	151	136	148	157	110	162
y	191	288	410	482	513	506	468	477	541

9.9 The total monthly expenditure (E) for family size (N) for five households is as follows:

E	250	300	410	450	565
N	2	3	4	5	6

Find regression equation of E on N. What is likely to be the expenditure for a household of size 8?

9.10 The following table gives the value of exports of raw cotton from India and the value of imports of manufactured cotton goods into India during the year 1990-91 to 1996-97:

Year	Exports of raw cotton	Import of manufactured cotton
1990-91	42	56
1991-92	44	49
1992-93	58	53
1993-94	55	58
1994-95	89	65
1995-96	98	74
1996-97	69	58

Calculate the coefficient of correlation between the values of exports and imports and obtain the equation of regression lines.

9.11 Calculate two regression lines from the following table giving the marks obtained in Economics and Statistics:

x/y	4-8	8-12	12-16	16-20	Total
8-14	11	6	2	1	20
14-20	5	12	15	8	40
20-26	-	2	3	15	20
Total	16	20	20	24	80

9.12 Obtain the regression equations of y on x and x on y:

x/y	5-15	15-25	25-35	35-45
0-10	1	1	-	-
10-20	3	6	5	1
20-30	1	8	9	2
30-40	-	3	9	3
40-50	-	-	4	4

9.13 For two series x and y which are correlated, the lines of regression of y on x and x on y are respectively $y = x + 5$ and $16x = 9y + 95$, also $\sigma_y = 4$, find σ_x and r_{xy}.

9.14 For 50 students of a class, regression equation of marks in Statistics (x) on marks in Accountancy (y) in $3y - 5x + 180 = 0$. The mean marks in Accountancy is 44 and the variance of marks in Statistics is $\frac{9}{16}$ of the variance of marks in Accountancy. Find the mean marks in Statistics and coefficient of correlation between marks in two subjects.

9.15 Given that: Regression equations: $x + 2y - 5 = 0$, $2x + 3y - 8 = 0$ and $\sigma_x^2 = 12$. Find: (i) $\bar{x}, \bar{y}$, (ii) σ_y^2, (iii) r_{xy}

9.16 The equation of two regression lines are as follows: $4x - 5y + 30 = 0$, $20x - 9y - 107 = 0$, which one of these is the line of regression of 'y on x'? and why?

9.17 Two regression equations are as follows: $3x + 2y - 26 = 0, 6x + y - 31 = 0$. Find which is that '$y$ on x' and that of 'x on y'. Find the mean of x and y the correlation coefficient between them. Also find out the ratio of the standard deviations of the two variables.

9.18 The equation of regression line between two variables are as follows: $2x - 3y = 0$ and $4y - 5x - 8 = 0$. Find $\bar{x}$and $\bar{y}$, both regression coefficient and correlation coefficient between x and y.

9.19 In a partially destroyed record the following data are available: Variance of $x = 25$, Regression equation of x on y, $5x - y = 22$ and Regression equation of y on x, $64x - 45y = 24$. Find: (i) Arithmetic mean of x (ii) Standard deviation of y (iii) Regression coefficient of y on x (iv) Correlation coefficient between x nad y.

9.20 For a bivariate data, the mean value of $x = 52.5$, the mean value of $y = 30.5$, the regression coefficient of y on $x = -1.6$ and correlation coefficient between x and y is -0.8. Find (i) regression coefficient of x on y and (ii) the most likely value of y when $x = 50$.

9.21 For 10 observation on price (x) and supply (y) the following data was obtained (in appropriate units):

$$\Sigma x = 130, \Sigma y = 220, \Sigma x^2 = 2,288, \Sigma y^2 = 5,506, \Sigma xy = 3,467$$

Obtain the line of regression of y on x and estimate the supply when the price is 16 units.

9.2 MULTIPLE REGRESSION

Multiple regression is the extension of simple regression. In multiple regression we take into account the effect of more than one independent variable on the dependent variable.

Multiple regression is the appropriate technique when we want to investigate the effects of several variables simultaneously on a dependent variable.

9.3 MULTIPLE REGRESSION EQUATION

Let Y = dependent variable, and $x_1, x_2, x_3, ...$, are independent variables, then multiple regression equation of Y on $x_1, x_2, x_3, ...$, is written as

$$y_c = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + ...$$

where y_c = computed value of dependent variable y and $b_1, b_2, b_3, ...$ are multiple regression coefficients.

The multiple regression equation expresses the average relationship of the variables and on the basis of this average relationship, an estimate of dependent variable is made. The multiple regression equation shows the simultaneous effect of a number of independent variables.

9.3.1 Specific Example:

Let x_1 = yield of wheat in 10 hectare land
x_2 = amount of rainfall
x_3 = amount of fertilizer used.

Then multiple regression equation x_1 on x_2 and x_3 is

$$x_1 = a + b_2 x_2 + b_3 x_3$$

or

$$x_1 = a_{1.23} + b_{12.3} x_2 + b_{13.2} x_3$$

where $a_{1.23}$ = constant, $b_{12.3}$ = partial regression coefficient which measures the change in x_1 for an unit change in x_2, keeping x_3 as constant, and $b_{13.2}$ = Partial regression coefficient which measures the change in x_1 for an unit change in x_3, keeping x_2 as constant.

Remark: When x_1, x_2, x_3 are deviations from their respective means, then multiple regression equation is $x_1 = b_{12.3} x_2 + b_{13.2} x_3$

9.4 COMPUTATION OF PARTIAL REGRESSION COEFFICIENTS

$$b_{12.3} = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \times \frac{\sigma_1}{\sigma_2}$$

$$b_{13.2} = \frac{r_{13} - r_{12} r_{32}}{1 - r_{32}^2} \times \frac{\sigma_1}{\sigma_3}$$

or

$$b_{12.3} = r_{12.3} \times \frac{\sigma_{1.23}}{\sigma_{2.13}}$$

$$b_{13.2} = r_{13.2} \times \frac{\sigma_{1.32}}{\sigma_{3.12}}$$

Here: $b_{12.3}$ and $b_{13.2}$ are partial regression coefficients, σ_1, σ_2 and σ_3 are standard deviations of x_1, x_2 and x_3.

After finding the values of $b_{12.3}$ and $b_{13.2}$, the value of $a_{1.23}$ in obtained as follows:

$$a_{1.23} = \bar{x_1} - b_{12.3}\bar{x_2} - b_{13.2}\bar{x_3}$$

The estimated value of x_1 on x_2 and x_3 is denoted as $x_1.23$

9.5 RELATIONSHIP BETWEEN PARTIAL CORRELATION COEFFICIENT AND PARTIAL REGRESSION COEFFICIENT

As r_{12} is the square root of the regression coefficients b_{12} and b_{21} i.e., $r_{12} = \sqrt{b_{12} \times b_{21}}$ the partial correlation coefficient $r_{12.3}$ is the square root of the product of two partial regression coefficients:

$$r_{12.3} = \sqrt{b_{12.3} \times b_{21.3}}$$

or

$$r_{12.3}^2 = b_{12.3} \times b_{21.3}$$

Proof:

$$b_{12.3} \times b_{21.3} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \times \frac{\sigma_1}{\sigma_2}\right) \times \left(\frac{r_{21} - r_{23}r_{13}}{1 - r_{13}^2} \times \frac{\sigma_2}{\sigma_1}\right)$$

$$= \frac{(r_{12} - r_{13}r_{23})^2}{[1 - r_{23}^2][1 - r_{13}^2]} = r_{12.3}^2$$

■ **EXAMPLE 9.9**

Given: $r_{12} = 0.3, r_{13} = 0.5, r_{23} = 0.4$ and $\sigma_1 = 3, \sigma_2 = 4, \sigma_3 = 5$. Find the regression equations of x_1 on x_2 and x_3 where x_1, x_2 and x_3 have been measured from their actual means.

Solution:

Since x_1, x_2, x_3 are measured from their actual means, we have the multiple regression equation of x_1 on x_2 and x_3:

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3$$

Now,

$$b_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \times \frac{\sigma_1}{\sigma_2} = \frac{0.3 - (0.5 \times 0.4)}{1 - (0.4)^2} \times \frac{3}{4} = 0.9$$

$$b_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \times \frac{\sigma_1}{\sigma_2} = \frac{0.5 - (0.3)(0.4)}{1 - (0.4)^2} \times \frac{3}{5} = 0.27142$$

Substituting these values in (1), we have required equation:

$$x_1 = 0.9x_2 + 0.27x_3$$

□

9.6 MULTIPLE REGRESSION EQUATION BY LEAST SQUARE METHOD

As the method of least squares is applied for simple regression equation, the same way we apply for multiple regression equation. Here we have to form as many equations as the number of unknown quantities. After finding the unknown quantities, they are substituted in the regression equation. For the multiple regression equation of x_1 on x_2 and x_3, we proceed as follows:

Original Equation: $x_1 = a_{1.23} + b_{12.3}x_2 + b_{13.2}x_3$

Normal Equations:

$$\sum x_1 = na_{1.23} + b_{12.3}\sum x_2 + b_{13.2}\sum x_3 \tag{9.3}$$

$$\sum x_1x_2 = a_{1.23}\sum x_2 + b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3 \tag{9.4}$$

$$\sum x_1x_3 = a_{1.23}\sum x_3 + b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 \tag{9.5}$$

Similarly, other multiple regression equations and their normal equations may be obtained.

EXAMPLE 9.10

From the following data estimates x_3, when $x_1 = 10$ and $x_2 = 3$.

x_1	2	3	4	5	6
x_2	6	8	4	3	3
x_3	10	6	12	16	8

Solution:

x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	x_1^2	x_2^2	x_3^2
2	6	10	12	20	60	4	36	100
3	8	6	24	18	48	9	64	36
4	4	12	16	48	48	16	16	144
5	3	16	15	80	48	25	9	256
6	3	8	18	48	24	36	9	64
$\sum x_1 =$ 20	$\sum x_2 =$ 24	$\sum x_3 =$ 52	$\sum x_1x_2 =$ 85	$\sum x_1x_3 =$ 214	$\sum x_2x_3 =$ 228	$\sum x_1^2 =$ 90	$\sum x_2^2 =$ 134	$\sum x_3^2 =$ 600

Original Equation:

$$x_3 = a_{3.12} + b_{31.2}x_1 + b_{32.1}x_2$$

Normal Equations:

$$\sum x_3 = Na_{3.12} + b_{31.2}\sum x_1 + b_{32.1}\sum x_2$$

$$\sum x_3x_1 = a_{3.12}\sum x_1 + b_{31.2}\sum x_1^2 + b_{32.1}\sum x_1x_2$$

$$\sum x_3x_1 = a_{3.12}\sum x_2 + b_{31.2}\sum x_1x_2 + b_{32.1}\sum x_2^2$$

Substituting the values of $\sum x_1$,etc.

$$52 = 5a_{3.12} + 20b_{31.2} + 24b_{32.1}$$

$$214 = 20a_{3.12} + 90b_{31.2} + 85b_{32.1}$$

$$228 = 24a_{3.12} + 85b_{31.2} + 134b_{32.1}$$

On solving these equations, we get

$$a_{3.12} = 28.5970$$

$$b_{32.1} = -2.2388$$
$$b_{31.2} = -1.8627$$

Hence, the required equation

$$x_3 = 28.5969 - 1.86268x_1 - 2.2388x_2$$

The value of $x_3$when $x_1 = 10$ and $x_2 = 3$

$$\therefore x_3 = 28.5969 - (1.86268 \times 10) - (2.2388 \times 3) = 3.2537$$

□

■ EXAMPLE 9.11

Given $\bar{x_1} = 48, \sigma_1 = 3, r_{12} = 0.6$, $\bar{x_2} = 40, \sigma_2 = 4, r_{13} = 0.7$, $\bar{x_3} = 62, \sigma_3 = 5, r_{23} = 0.8$, estimate the value of x_3 when $x_1 = 30$ and $x_2 = 50$.

Solution:

Multiple Regression of x_3 on x_1 and x_2 will give the value of x_3 for x_1 = 30 and $x_2 = 50$, this equation is.

$$x_3 = a_{3.12} + b_{31.2}x_1 + b_{32.1}x_2 \tag{9.6}$$

The values of regression coefficients:

$$b_{31.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2} \times \frac{\sigma_3}{\sigma_1}$$
$$= \frac{0.7 - (0.6)(0.8)}{1 - (0.6)^2} \times \frac{5}{3} = 0.57$$
$$b_{32.1} = \frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2} \times \frac{\sigma_3}{\sigma_2}$$
$$= \frac{0.8 - (0.6)(0.7)}{1 - (0.6)^2} \times \frac{5}{4} = 0.742$$

Now

$$a_{3.12} = \bar{x_3} - b_{31.2}\bar{x_1} - b_{32.1}\bar{x_2}$$
$$= 62 - (0.57 \times 48) - (0.74 \times 40) = 5.04$$

Substitute the value of $a_{3.12}$ in equation 9.6

$$x_3 = 5.04 + (0.57 \times 30) + (0.74 \times 50) = 59.14$$

$\therefore$ Thus estimated value of $x_3 = 59.14$

□

Problems

9.22 If $r_{12} = 0.28, r_{23} = 0.49$ and $r_{13} = 0.51$ calculate $r_{12.3}$ and $r_{13.2}$

9.23 Given $r_{12} = 0.28$, $r_{23} = 0.49$, $r_{13} = 0.51$ $\sigma_1 = 2.7$, $\sigma_2 = 2.4$, $\sigma_3 = 0.27$ Find the regression equation of x_3 on x_1 and x_2. If the variables have been measured from their actual means.

9.24 From the following data, find regression equation:

Wheat yield (x_1)(per hectare Quintals)	40	45	50	65	70	70	80
Use of Fertilizers (x_2)(Kg. per hectare)	10	20	30	40	50	60	70
Rainfall (x_3) (inches)	36	33	37	37	34	32	36

Also calculate the value of x_3 when $x_1 = 45$ and $x_1 = 30$

9.25 From the data given below, estimate the value of x_3 when $x_1 = 58$ and $x_2 = 52.5$

$$\bar{x_1} = 55.95\ \sigma_1 = 2.26,\ r_{12} = 0.578$$

$$\bar{x_2} = 51.48\ \sigma_2 = 4.39,\ r_{13} = 0.581$$

$$\bar{x_3} = 56.03\ \sigma_3 = 4.41,\ r_{23} = 0.974$$

9.26 Find the multiple linear regression of x_1 on x_2 and x_3 from the data relating to three variables given below:

x_1	4	6	7	9	13	15
x_2	15	12	8	6	4	3
x_3	30	24	20	14	10	4

CHAPTER 10

THEORY OF PROBABILITY

Probability theory began in seventeenth century in France when the two great French mathematicians, Blaise Pascal and Pierre De Fermat, corresponded over two problems from games of chance. Problems like those Pascal and Fermat solved continued to influence such early researchers as Huygens, Bernoulli, and De Moivre in establishing a mathematical theory of probability. Today, probability theory is a well established branch of mathematics that finds applications in every area of scholarly activity from music to engineering, and in daily experience from weather prediction to predicting the risks of new medical treatments.

10.1 BASIC CONCEPTS

10.1.1 Experiments

An experiment repeated under essentially homogeneous and similar conditions results in an outcome, which is unique or not unique but may be one of the several possible outcomes. When the result is unique then the experiment is called a *deterministic experiment*. Any experiment whose outcome cannot be predicted in advance, but is one of the set of possible outcomes, is called a *random experiment*. If any experiment as being performed repeatedly, each repetition is called a trial. We observe an outcome for each trial.

If a coin is tossed, the result of a particular tossing, i.e., whether it will fall with head or tail can not be predicted. A die with faces marked with numbers numbers 1,2,3,4,5,6. When we throw a die the number on upper face, we can not predict which face will be on the top.

10.1.2 Sample Space

The set of all possible outcomes of a random experiment is called the *sample space*. e.g., The sample space for tossing coins contains just two outcomes - Head or Tail. If we denote outcome of head as H and of tail as T , The sample space for the random experiment of tossing of coins is

$$S = \{H, T\}$$

Similarly, the sample space of throwing of a die will be

$$S = \{1, 2, 3, 4, 5, 6\}$$

Each possible outcome given in the sample space is called an element or sample point. Here in case of die, the sample points are 1,2,3,4,5,6.

Further sample space may be classified in two way based on number of sample points are finite or infinite. A sample space is called *discrete* if it contains finite or finitely many countably infinite sample points. for example-when a coin and die is thrown simultaneously, the sample space is given below is discrete:

$$S = \{(H,1),(H,2),(H,3),(H,4),(H,5),(H,6),(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}$$

In other example, a coin is tossed until a head turns up. The sample space is

$$S = \{H, TH, TTH, TTTH, ...\}$$

where the number of sample points is infinite but there is one to one correspondence between the sample points as natural numbers, i.e. S is countable infinite. Hence is a discrete sample space.

On the other hand a sample space is said to be *continuous* if it contains uncountable number of sample points. e.g., All points on a straight line, All points inside a circle of radius 4, etc.

10.1.3 Events

Any subset of the sample space S, associated with any random experiment is called an *event*. For example when a die is thrown, then the sample space is

$$S = 1, 2, 3, 4, 5, 6$$

and its subset

$$E_1 = \{2, 4, 6\}$$

represents an event, which represents ‘The number on upper face is even’.

Since a set is also a subset of itself, we say that sample space is itself an event and call it a *sure event*. An event that contains no sample point is called an *impossible event*. An event containing exactly on sample point is called *simple or elementary event*.

During the performance of an experiment, those point which entail or assume the occurrence of an event A, are called *favorable* to that event. If any of the sample point favorable to an event A occurs in a trial, we say that the events occurs at this trial.

10.1.4 Complex or Composite Event

The union of simple events is called *composite event.* In other words if an event can be decomposed into simple events, then it is called a composite event. Let A and B be two simple events associated with sample space S, then the event which consists of all the sample points which belong to A or B or both, is called the union of A and B as $A \cup B$ or $A + B$. No sample point is taken twice in $A \cup B$. For example, when three coins are tossed together, then the event E=at least two heads, is a composite event.

Here,

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$
$$A(\text{Event of two heads}) = \{HHT, HTH, THH\}$$
$$B(\text{Event of Three heads}) = \{HHH\}$$
$$E = A \cup B = \{HHH, HHT, HTH, THH\}$$

10.1.5 Compound or Joint Event

The occurrence of two or more events together is called a *compound event*. In other words the intersection of two or more events is called compound event. Let A and B be any two simple events associated with a sample space S, then the intersection of A and B, written as $A \cap B$ or AB, which contains all those point of S, which are common in A and B. e.g.,

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$A(\text{Event of Even points}) = \{2, 4, 6\}$$
$$B(\text{Event of multiple of Three}) = \{3, 6\}$$
$$E = A \cap B = \{6\}$$

here E is compound event.

10.1.6 Mutually exclusive or disjoint events

Two events are called *mutually exclusive* if they can not occur simultaneously. In other words two or more events are said to be mutually exclusive if the occurrence of any one of the event precludes the occurrence of the other events. There are no sample points are common in mutually exclusive events. For example, if we toss a coin once, the events of head and tail are mutually exclusive because both can not occur same time.

10.2 ADDITION THEOREM OF PROBABILITY

If $A_1, A_2, \ldots, A_k$ be k mutually exclusive events on a sample space associated with a random experiment, then the probability of happening one of them is the sum of their individual probabilities. In symbols,

$$P(A_1 + A_2 + \ldots + A_k) = P(A_1) + P(A_2) + \ldots + P(A_k)$$

or

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \ldots + P(A_k)$$

Proof. Let total ways = N. Let $a_1, a_2, \ldots, a_k$ be the favorable ways to the events $A_1, A_2, \ldots, A_k$ respectively. Then the number of favorable ways to any one of the events

$$= a_1 + a_2 + \ldots + a_k$$

$$\begin{aligned}
\therefore P(A_1 \cup A_2 \cup \ldots \cup A_k) &= \frac{a_1 + a_2 + \ldots + a_k}{N} \\
&= \frac{a_1}{N} + \frac{a_2}{N} + \ldots + \frac{a_k}{N} \\
&= P(A_1) + P(A_2) + \ldots + P(A_k)
\end{aligned}$$

□

■ **EXAMPLE 10.1**

From 30 tickets marked with the first 30 numerals one is drawn at random. Find the probability that the number on this ticket is a multiple of 3 or 11.

Solution: The multiples of 3 are: 3, 6, 9, 12, 15, 18, 21, 24, 27, and 30, which are 10. Therefore, by definition probability that the number on the ticket is a multiple of 3

$$P(A) = \frac{10}{30}$$

The multiples of 11 are: 11, 22. i.e. whose number is 2. Probability that the number on the ticket is a multiple of 11

$$P(B) = \frac{2}{30}$$

$\therefore$ Required probability

$$P(A \text{ or } B) = P(A) + P(B) = \frac{10}{30} + \frac{2}{30} = \frac{12}{30} = \frac{2}{5}$$

□

■ **EXAMPLE 10.2**

From a pack of 52 cards one card is drawn at random. What is the probability that it will be a queen of clubs or king of diamonds?

Solution:

The probability of drawing a queen of clubs $= \frac{1}{52}$ and the probability of drawing a king of diamonds $= \frac{1}{52}$

$\therefore$ Required probability

$$= \frac{1}{52} + \frac{1}{52} = \frac{2}{52} = \frac{1}{26}$$

because the two events are mutually exclusive.

□

■ **EXAMPLE 10.3**

Two cards are randomly drawn from a pack of 52 cards and thrown away. What is the probability of drawing an ace in a simple draw from the remaining 50 cards?

Solution:

There are three cases when two cards have been drawn from a pack of 52 cards:

Case 1. There is no ace in cards thrown away, the probability of getting an ace out of 50

$$= \frac{4}{50}$$

Case 2. There is one ace in the cards thrown away, the probability of getting an ace out of 50

$$= \frac{3}{50}$$

Case 3. there are two aces in the cards thrown away, the probability of getting an ace out of 50

$$= \frac{2}{50}$$

Hence, Required probability

$$= \frac{4}{50} + \frac{3}{50} + \frac{2}{50} = \frac{9}{50}$$

□

■ EXAMPLE 10.4

A bag contains 5 white, 4 black, 3 yellow and 4 red balls. What is the probability of getting a black or red ball at random in a single draw of one?

Solution:

Total balls $= 5+4+3+2 = 16$

$$\text{Probability of getting a black ball} = \tfrac{4}{16}$$

$$\text{Probability of getting a red ball} = \tfrac{4}{16}$$

$\therefore$ Probability of getting a black or a red ball

$$= \frac{4}{16} + \frac{4}{16} = \frac{4+4}{16} = \frac{8}{16} = \frac{1}{2} = 0.5$$

□

■ EXAMPLE 10.5

What is the probability of getting a total of either 7 or 11 in a single throw with two dice?

Solution:

Total ways, $N = 36$. Let A = the sum is 7 which can be obtained as (1,6),(2,5),(3,4),(4,3),(5,2),(6,1). Favorable ways to $A = 6$

$$\therefore P(A) = \frac{6}{36}$$

Let B = The sum of 11, which can be obtained as (5,6),(6,5). Favorable ways to $B = 2$

$$\therefore P(B) = \frac{2}{36}$$

$$\therefore P(7 or 11) = P(A \cup B)$$

$$= P(A) + P(B) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} = \frac{2}{9}$$

□

10.3 INCLUSION -EXCLUSION FORMULA

IF A and B are any two events then the probability of happening at least one is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

or

$$P(A+B) = P(A) + P(B) - P(AB)$$

Proof. If A and B are any two events, then the number of favorable ways for happening of at least one of the these events, then favorable ways to $A \cup B$ are

$$m_1 + m_2 - m_3$$

where m_1= favorable ways to A, m_2= favorable ways to B, m_3 = favorable ways to $A \cap B$

$$\therefore P(A \cup B) = \frac{m_1 + m_2 - m_3}{n} = \frac{m_1}{n} + \frac{m_2}{n} - \frac{m_3}{n}$$

$$= P(A) + P(B) - P(A \cap B)$$

□

Remark :If A, B, C are any these events, then the probability of happening at least one of them is given by
$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

10.3.1 Particular Formulas

If A and B are any two events

1. $P(A) = P(A \cap B) + P(A \cap B^c)$
2. $P(B) = P(A \cap B) + P(A^c \cap B)$
3. Probability of happening at exactly one of them

$$\begin{aligned} P(A \cap B^c or A^c \cap B) &= P(A \cap B^c) + P(A^c \cap B) \\ &= P(A) + P(B) - 2P(A \cap B) \end{aligned}$$

4. Probability of happening at least one of them

$$\begin{aligned} P(A or B) &= P(A \cap B^c) + P(A^c \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \\ &= 1 - P(A^c \cap B^c) \\ &= 1 - P(\text{None of the events occur}) \end{aligned}$$

■ **EXAMPLE 10.6**

What is the probability of drawing a card of heart or an ace from a pack of cards?

Solution:
There are 52 cards in a pack ; 13 are of heart, 4 are aces, 1 is an ace of heart.
Hence,
Probability of drawing a card of heart,

$$P(A) = \frac{13}{52}$$

Probability of drawing a card of ace,

$$P(B) = \frac{4}{52}$$

Probability of drawing an ace of heart,

$$P(A\cap B)=\frac{1}{52}$$

Hence, probability of drawing a card of heart or an ace

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

$$=\frac{13}{52}+\frac{4}{52}-\frac{1}{52}=\frac{16}{52}=\frac{4}{13}$$

□

■ EXAMPLE 10.7

What is the probability of getting either a doublet or total 4 in a single throw with two dice?

Solution:

Total ways, N = 36, Let Event of doublets

$$A=\{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$$

and event of sum 4:

$$B=\{(1,3),(2,2),(3,1)\}$$

Here it is obvious that $P(A\cap B)=0$. Then $\therefore$ Required Probability

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)=\frac{6}{36}+\frac{3}{36}-\frac{1}{36}=\frac{8}{36}=\frac{2}{9}$$

□

■ EXAMPLE 10.8

Find the chance that a leap year selected at will contain (i) 53 Sunday and 53 Monday (ii) 53 Sunday or 53 Monday?

Solution:

A leap year consists of 366 days and therefore continuous 52 complete weeks and 2 days over. These two days may have the following 7 pairs :

(1) Monday and Tuesday
(2) Tuesday and Wednesday
(3) Wednesday and Thursday
(4) Thursday and Friday
(5) Friday and Saturday
(6) Saturday and Sunday
(7) Sunday and Monday

(i) 53 Sunday and 53 Monday are in one way only. Hence Required probability $=\frac{1}{7}$
(ii)$A=53$ Sunday: 2 ways, $B=53$ Monday: 2 ways, $A\cap B=53$ Sunday and 53 Monday: 1 way

$$\therefore P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

$$=\frac{2}{7}+\frac{2}{7}-\frac{1}{7}=\frac{3}{7}$$

□

■ EXAMPLE 10.9

There are 10boys and 20 girls in a class; in which half boys and half girls have blue eyes. One representative is selected at random from the class. What is the probability that he is a boy or his eyes are blue?

Solution:
The given information in tabular form

	Boys	Girls	Total
Blue eyes	5	10	15
Not blue eyes	5	10	15
Total	10	20	30

Let A = Boy

$$\therefore P(A) = \frac{10}{30}$$

B = Blue eyes

$$\therefore P(B) = \frac{15}{30}$$

$A \cap B =$ A boy with blue eyes

$$\therefore P(A \cap B) = \frac{5}{30}$$

Hence Required Probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{10}{30} + \frac{15}{30} - \frac{5}{30} = \frac{20}{30} = \frac{2}{3}$$

□

10.4 CONDITIONAL PROBABILITY

The probability that the event B will occur, it being known that A has occurred is called the conditional probability of B and is denoted by $P(B/A)$.In symbols,

$$\text{Conditional probability of } B \text{ when } A \text{ has happened} = P(B/A) = \frac{P(A \cap B)}{P(A)}, P(A) \neq 0$$

$$\Rightarrow P(A \cap B) = P(A)P(B/A)$$

■ EXAMPLE 10.10

Out of a well shuffled pack os cards a card is drawn, which is found a king. Another card is drawn without replacing it.Find the probability of second card being a king. Also find the probability of both cards being kings.

Solution:
Let A: a king is drawn in first draw

$$\therefore P(A) = \frac{4}{52}$$

Let B: a king is drawn in second draw.
There are 3 kings now, and total number of cards is 51. Hence

$$= P(B/A) = \frac{3}{51}$$

The probability that both cards are king

$$P(A \cap B) = P(A)P(B/A) = \frac{4}{52} \times \frac{3}{51} = \frac{1}{221}$$

□

■ EXAMPLE 10.11

The probability that a person stopping at a petrol pump will ask to have his tyres checked is 0.12; the probability that he will ask to have his oil checked 0.29 and the probability that he will ask to have both of them checked is 0.07. Find the probability that

(i) a person who will check oil, will also have tyres checked.
(ii) a person stopping at the petrol pump will have either tyres or oil checked.
(iii) a person stooping at the petrol pump will have neither his tyres nor his oil checked.

Solution:

Let A = To check tyre, B = To check oil, then

$$P(A) = 0.12,\ P(B) = 0.29,\ P(A \cap B) = 0.07$$

(i) $P(A/B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.07}{0.29} = \dfrac{7}{29} = 0.24$
(ii) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.12 + 0.29 - 0.07 = 0.34$
(iii)$P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B) = 1 - 0.34 = 0.66$

□

10.5 DEPENDENT AND INDEPENDENT EVENTS

When probability of happening event B is affected by the happening of the event A, then event B is said to be dependent on the event A.When the probability of happening event B is not affected by the happening of the event A, then event B is said to be independent of A. In symbols, If $P(B/A) = P(B), B$ is independent of A. If $P(A/B) = P(A), A$ is independent of B. Thus two events A and B are said to be independent if and only if

$$P(A \cap B) = P(A)P(B)$$

For example,

(i) A coin is tossed twice. Then A = head on first toss and B = tail on second toss are independent.

(ii) If a king is drawn from a pack of cards and is replaced, then the probability of drawing a king second time $P(B/A) = P(B) = \frac{4}{52}$; then A and B are independent.

10.6 MULTIPLICATION THEOREM OF PROBABILITY

1. The probability of happening of the two independent events A and B together is equal to the product of their individual probabilities. That is

$$P(A \cap B = P(A)P(B)$$

2. If A and B are not independent, then the probability $P(A \cap B)$ of their simultaneous occurrence is equal to the product of the probability of $A, P(A)$ and the conditional probability $P(B/A)$.In symbols
$$P(A \cap B) = P(A)P(B/A)$$
or
$$P(A \cap B) = P(B)P(A/B)$$

10.6.1 Specific Formulas

1. If $P(A \cap B) = P(A)P(B)$, then
$$P(A \cap B^c) = P(A)P(B^c)$$
$$P(A^c + B) = P(A^c)P(B), P(A^c \cap B^c) = P(A^c)P(B^c)$$

2. If the events A and B are independent
$$\begin{aligned} P(AorB) &= P(A) + P(B) - P(A)P(B) \\ &= 1 - P(\text{neither } A \text{ or } B) \\ &= 1 - P(A^c \cap B^c) \\ &= 1 - P(A^c)P(B^c) \end{aligned}$$

3. If the events A and B are independent
$$\begin{aligned} P(A \cap B^c) + P(A^c \cap B) &= P(\text{Exactly one}) \\ &= P(A)P(B^c) + P(A^c)P(B) \end{aligned}$$

4. If the events A, B, C, D are independent
$$P(ABCD) = P(A)P(B)P(C)P(D)$$
$$\begin{aligned} P(AorBorCorD) &= 1 - P(A^c \cap B^c \cap C^c \cap D^c) \\ &= 1 - P(A^c)P(B^c)P(C^c)P(D^c) \end{aligned}$$

■ **EXAMPLE 10.12**

The probability of the occurrence of two independent events A and B are 0.7 and 0.54 respectively. What is the probability that only one the A or B will occur in an experiment?

Solution:

$$\begin{aligned} \text{P(One and only one)} &= P(A \cap \bar{B}) + P(\bar{A} \cap B) \\ &= P(A)P(\bar{B}) + P(\bar{A})P(B) \\ &(\because A \text{ and } B \text{ are independent }) \\ &= P(A)[1 - P(B)] + [1 - P(A)]P(B) \\ &= 0.7(1 - 0.54) + (1 - 0.7)(0.54) = 0.484 \end{aligned}$$

Alternative Method .

$$\begin{aligned} \text{P(One and only one)} &= P(A)+P(B)-2P(A\cap B) \\ &= P(A)+P(B)-2P(A)P(B) \\ &= 0.7+0.54-2\times 0.7\times 0.54 = 0.484 \end{aligned}$$

□

■ EXAMPLE 10.13

A dice is thrown twice. (i) What is the probability that it falls with number 6 upwards in the first throw and an odd number upwards in second throw? What is the probability of getting a four both times?

Solution:

(i) P (a 6 in first throw),

$$P(A) = \frac{1}{6}$$

P (an odd number in second throw),

$$P(B) = \frac{3}{6}$$

$\therefore$Required Probability

$$P(A\cap B) = P(A)P(B) = \frac{3}{36} = \frac{1}{12}$$

(ii) P (a 4 in first throw)

$$P(A) = \frac{1}{6}$$

P(a 4 in second throw)

$$P(B) = \frac{1}{6}$$

$\therefore$P(a 4 both times)

$$P(A\cap B) = P(A)P(B) = \frac{1}{6}\times\frac{1}{6} = \frac{1}{36}$$

□

■ EXAMPLE 10.14

A University has to select a examiner from a list of 50 persons 20 of them are women and 30 are men; 10 of them know Hindi and 40 do not;15 of them are teachers and remaining 35 are not.What is the probability that university will select a Hindi knowing women teacher ?

Solution:

Let A = Selection of Hindi knowing person
B = Selection of a women
C = Selection of a teacher

$$\because P(A) = \frac{10}{50}, P(B) = \frac{20}{50}, P(C) = \frac{15}{50}$$

$\therefore$ Required Probability(assuming independence)

$$P(A\cap B\cap C) = P(A)P(B)P(C) = \frac{10}{50}\times\frac{20}{50}\times\frac{15}{50} = \frac{3}{125}$$

□

EXAMPLE 10.15

A bag contains 14 balls of which 6 are red, 3 yellow and 5 black . 3 balls are drawn successively without replacement. What is the probability that they are drawn in order red, yellow,black?

Solution:

Let the event under study, A = $A_1A_2A_3$
The probability of drawing a red ball out of 14 balls,

$$P(A_1) = \frac{6}{14}$$

The probability of drawing a yellow ball out of 13 balls,

$$P(A_2) = \frac{3}{13}$$

The probability of drawing a black ball out of 12 balls,

$$P(A_3) = \frac{5}{12}$$

∴ Required Probability

$$P(A_1)P(A_2)P(A_3) = \frac{6}{14} \times \frac{3}{13} \times \frac{5}{12} = \frac{15}{364}$$

□

EXAMPLE 10.16

The odds against A solving a sum are 7 : 6 and the odds in favor of B solving the same are 11: 8. What is the probability that the sum will be solved if both of them try it?

Solution:

Probability that A solves the question $= \frac{6}{7+6} = \frac{6}{13}$
Probability that A does not solve the question $= 1 - \frac{6}{13} = \frac{7}{13}$
Probability of that B solves the question = $\frac{11}{11+8} = \frac{11}{19}$
Probability that B does not solve the question $= 1 - \frac{11}{19} = \frac{8}{19}$
Probability that both do not solve the question $= (1 - \frac{16}{13})(1 - \frac{11}{19})$
Hence, Probability that is question is solved $= 1 - (1 - \frac{6}{13})(1 - \frac{11}{19})$

$$= 1 - \frac{7}{13} \times \frac{8}{19} = 0.773$$

□

EXAMPLE 10.17

A speaks truth in 80 % cases and B in 90 % cases. In what percentage of cases will they contradict each other in stating a fact?

Solution:

Let A = A speaks truth, $P(A) = \frac{80}{100} = 0.8$
B = B speaks truth, $P(B) = \frac{90}{100} = 0.9$
They will contradict each other $= (A \cap \bar{B}) or (\bar{A} \cap B)$
∴ Required probability $= P(A \cap \bar{B}) + P(\bar{A} \cap B)$

$$= P(A)P(B) + P(\bar{A})P(B)$$

$$= 0.8 \times (1 - 0.9) + (1 - 0.8) \times 0.9 = 0.26$$

that is they contradict each other in stating a fact in 26 % cases .

□

■ **EXAMPLE 10.18**

A can solve 75 % of the problems in statistics and B can solve 70 %. What is the probability that either A or B can solve a problem chosen at random?

Solution:

Such problems can be solved in three ways:
Let A = A can solve the problem, B = B can solve the problem, Thus
$A \cap B = A$ and B both can solve the problem
$A \cap \bar{B} = A$ can solve the problem but B not
$\bar{A} \cap B$ = B can solve the problem but A not
$A \cap B =$ Neither A nor B can solve the problem
$A \cup B$ = A or B can solve the problem

Method I.

$$P(A \cup B) = P(A) + P(B) - P(A) \times P(B)$$

$$= \frac{75}{100} + \frac{70}{100} - \frac{75}{100} \times \frac{70}{100} = \frac{37}{40}$$

Method II.

$$P(A \cup B) = P(A \cap \bar{B}) + P(\bar{A} \cup B) + P(A \cap B)$$

$$= \frac{3}{4} \times \frac{3}{10} + \frac{1}{4} \times \frac{7}{10} + \frac{3}{4} \times \frac{7}{10} = \frac{37}{40}$$

Method III.

$$P(A \cup B) = 1 - P(\bar{A} \cap \bar{B})$$

$$= 1 - P(\bar{A})P(\bar{B})$$

$$= 1 - \frac{1}{4} \times \frac{3}{10} = \frac{37}{40}$$

□

■ **EXAMPLE 10.19**

A can hit a target 3 times in 5 shots, B 2 times in 5 shots and C 3 times in 4 shots. They fire a volley. What is the probability that (i) two shots hit the target? (ii) at least two shots hit the target.

Solution: Let A = A hits the target,

$$P(A) = \frac{3}{5},\ P(\bar{A}) = \frac{2}{5}$$

B = B hits the target,

$$P(B) = \frac{2}{5},\ P(\bar{B}) = \frac{3}{5}$$

C = C hits the target,

$$P(C) = \frac{3}{4},\ P(\bar{C}) = \frac{1}{4}$$

Two shots can hit in there mutually exclusive ways:
(a) A and B hit C not, $A \cap B \cap \bar{C}$
(b) A and C hit B not, $A \cap \bar{B} \cap C$
(c) B and C hit A not, $\bar{A} \cap B \cap C$
(i) Required Probability

$$\begin{aligned}
&= P(A \cap B \cap \bar{C}) + P(A \cap \bar{B} \cap C) + P(\bar{A} + B + C) \\
&= P(A)P(B)P(\bar{C}) + P(A)P(\bar{B})P(C) + P(\bar{A})P(B)P(C) \\
&= \frac{3}{5} \times \frac{2}{5} \times \frac{1}{4} + \frac{3}{5} \times \frac{3}{5} \times \frac{3}{4} + \frac{2}{5} \times \frac{2}{5} \times \frac{3}{4} = \frac{9}{20}
\end{aligned}$$

(ii) Required Probability

$$\begin{aligned}
&= P(A \cap B \cap \bar{C}) + P(A \cap \bar{B} \cap C) + P(\bar{A} \cap B \cap C) + P(A \cap B \cap C) \\
&= \frac{6}{100} + \frac{27}{100} + \frac{12}{100} + \frac{18}{100} = \frac{63}{100} = 0.63
\end{aligned}$$

□

■ EXAMPLE 10.20

Three groups of children contain 3 girls and 1 boy, 2 girls and 2 boys, 1 girl and 3 boys. One child is selected from each group. Find the chance that the three selected consist of 1 girl and 2 boys.

Solution:

Group	I	II	III
Number of Children	3 girls 1 boy	2girls 2 boys	1 girls 3 boys

1 girl and 2 boys can be selected as follows:

A_1 = girl from 1st group, boy from 2nd, boy from 3rd

A_2 =boy from 1st group, girl from 2nd, boy from 3rd

A_3 =boy from 1st group, boy from 2nd, girl from 3rd

$$\therefore P(A_1) = \frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{18}{64}$$

$$P(A_2) = \frac{1}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{6}{64}$$

$$P(A_3) = \frac{1}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{2}{64}$$

$\therefore$ Required Probability $= P$ (1 girl and 2 boys)

$$= P(A_1) + P(A_2) + P(A_3) = \frac{18}{64} + \frac{6}{64} + \frac{2}{64} = \frac{26}{64} = \frac{13}{32}$$

□

EXAMPLE 10.21

Two urns A and B contains 2 white and 5 black balls and 3 white and 6 black balls respectively. A ball is drawn from urn A and put into urn B and then a ball is drawn from urn B. Find the probability of this ball being black.

Solution:

Let A_1 = To put a white ball in urn B from urn A, A_2 = To put a black ball in urn B from urn A, B = To draw a black ball from urn B. Hence

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) \\ &= P(A_1)P(B/A_1) + P(A_2)P(B/A_2) \\ &= \frac{2}{7} \times \frac{6}{10} + \frac{5}{7} \times \frac{7}{10} = \frac{47}{70} \end{aligned}$$

□

Problems

10.1 A card is drawn at random from a well shuffled pack of cards. Find the probability that the card is (i) a court card, (ii) a card of space, (iii) an ace.

10.2 A card is drawn at random from a pack of playing cards. What is the probability that the card is (i) a knave, (ii) of red color (iii) the king of diamond.

10.3 A die is thrown, Find the probability (i) of getting a 4 or 6 and (ii) of getting more than 2 with an ordinary die.

10.4 In a question of Mathematics the students are asked to solve 4 questions out of 9, while in a question paper of Statistics students are asked the solve 5 questions and of 10. Which question paper got better choice and how?

10.5 What is the probability of selecting a king from a pinochle deck ? (A pinochle deck consists of 2 aces, 2kings, 2 queens,2 jacks, 2 tens and 2 nines of each suit. there are no other cards.)

10.6 Two dice are thrown simultaneously. Find the probability:

- **a)** of getting odd digit on first die
- **b)** of getting a sum of 9
- **c)** of getting a sum of 11
- **d)** of not throwing a total of 9.

10.7 In a sample of 120 radios the following information is available:

No of defects	0	1	2	3
No of radios	15	80	20	5

10.8 A bag contains 5 white, 4 black, 7 yellow and 6 red balls. What is the probability of getting a black or red ball at random is a single draw of one?

10.7 USE OF PERMUTATION AND COMBINATION IN PROBABILITY

EXAMPLE 10.22

There are 5 doors in a room. Four persons enter the room. Find the probability of their entering through different doors.

Solution:

First person can enter the room from any door out of 5, then second person can enter the room any door out of 5 and so on. Thus, Total ways,

$$n = 5 \times 5 \times 5 \times 5 = 5^4$$

Favorable ways,

$$a = {}^5P_4 = 5 \times 4 \times 3 \times 2$$

$\therefore$ Required Probability

$$= \frac{a}{n} = \frac{5 \times 4 \times 3 \times 2}{5 \times 5 \times 5 \times 5} = \frac{24}{125}$$

□

EXAMPLE 10.23

Out of 6 Indians and 4 Americans, a committee of five is constituted . Find the probability that there are exactly two Indians in it.

Solution:

Total number of persons, 6+4 =10

The number of ways of constituting a committee of 5 out of 10,

$$n = {}^{10}C_5 = \frac{10!}{5!5!} = \frac{10 \times 9 \times 8 \times 7 \times 6}{5 \times 4 \times 3 \times 2 \times 1} = 126$$

The number of ways of selecting 2 Indian out of 6

$$= {}^6C_2 = \frac{6 \times 5}{2 \times 1} = 15$$

The number of ways of selecting 3 American out of 4

$$= {}^4C_3 = \frac{4!}{3!(4-3)!} = 4$$

Favorable ways,

$$a = 15 \times 4 = 60$$

$\therefore$ Required Probability

$$= \frac{a}{n} = \frac{60}{126} = \frac{10}{21}$$

□

EXAMPLE 10.24

In a bag there are 4 red and 3 black balls. What is the probability of drawing the first ball red and second black, the third red and fourth black and so on; if they are drawn one at a time?

Solution:

Total balls = 4+3=7

The seven balls can be drawn in 7 ! ways. Thus total ways, n = 7!

Let the balls be in the order

R	B	R	B	R	B	R
1	2	3	4	5	6	7

The 4 red balls can be arranged in 4! ways at places 1, 3, 5, 7. The 3 black balls can be arranged in 3! ways at places 2, 4, 6. Therefore favorable ways = 4!3!

$\therefore$ Required Probability

$$= \frac{4!3!}{7!} = \frac{1}{35}$$

□

EXAMPLE 10.25

There are 4 Engineers and 3 Doctors in a firm. Three persons are put on duty at a time. What is the probability that there are 2 Engineers and 1 Doctors?

Solution:

Total number of persons $= 4+3 = 7$

Total number of ways of talking three persons out of 7 persons $=^7 C_3$

Favorable cases to the event, '2 Engineers and 1 Doctor' = ${}^4C_2 \times {}^3C_1$

$\therefore$ Required Probability

$$= \frac{{}^4C_2 \times {}^3C_1}{{}^7C_3} = \frac{18}{35}$$

□

EXAMPLE 10.26

There are 5 red and 8 black balls in a bag. Two successive draws of three balls each are made, the balls not being replaced after the first draw. What is the chance that in the first draw the balls were red and in the second black?

Solution: Probability of drawing 3 red balls in first draw,

$$P(A) = \frac{{}^5C_3}{{}^{13}C_3}$$

Probability of drawing 3 black balls in second draw,

$$P(B/A) = \frac{{}^8C_3}{{}^{10}C_3}$$

Hence, Required probability

$$P(A \cap B) = P(A)P(B/A) = \frac{{}^5C_3}{{}^{13}C_3} \times \frac{{}^8C_3}{{}^{10}C_3} = \frac{7}{429}$$

□

Problems

10.9 Two cities A and B are connected by five roads.If a person goes from city A to B by a road and returns, what is the probability that he does not return with the same road?

10.10 There are 5 hotels in a city. If 4 persons come to the city. find the probability that they will stay in different hotels.

10.11 There are 13 computers and 16 typing machines in a firm. Out of these two become out of order. Find the probability that they are of the same type.

10.12 In a bag 3 balls are red out of 10 balls. If the balls are drawn at a time from the bag, what is the probability that (i) at least one ball is of red color and (ii) both are of red color?

10.13 A bag contains 8 balls, identical except for color, of which 5 are red and 3 white. A man draws two balls at random. What is the probability that

(i) one of the balls drawn is white and the other red?

(ii) both are of the same color.

What would be the values of these probabilities if a ball is drawn, replaced and the another ball is drawn.

10.14 An urn contains 7 white and 3 black balls, 3 balls are drawn at random one by one with replacement. Find the probability that (i) at least one ball is black, and (ii) the first ball and the last ball are of different colors.

10.15 In a party of 12 persons, 7 are men and 5 are ladies. Two persons are selected at random. What is the probability of both being men?

10.16 A bag contains 6 white and 4 black balls and a second one 4 white and 8 black balls. One of the bag is chosen at random and a draw of 2 balls is made from it.Find the probability that one is white and the other is black.

10.17 there are 4 wholesalers and 5 retailers in a market.A salesman visits any three in a day. Find the probability that those three are wholesalers.

10.18 a bag contains 4 white, 5 red and 6 green balls, Three balls are drawn at random. What is the probability that a white, a red and a green ball are drawn?

10.8 BAYES THEOREM

Theorem 12. *An event A, can be affected by mutually exclusive and exhaustive events* $B_1, B_2, ..., B_n$*. If the prior probabilities* $P(B_1), P(B_2), ..., P(B_n)$ *are known and conditional probabilities* $P\left(\frac{A}{B_1}\right)$, $P\left(\frac{A}{B_2}\right)$, ..., $P\left(\frac{A}{B_n}\right)$ *are also known then a posterior probability* $p\left(\frac{B_n}{A}\right)$ *is given by*

$$P\left(\frac{B_n}{A}\right) = \frac{P(B_m)P(A/B_m)}{\sum_{m=1}^{n} P(B_m)P(A/B_m)}$$

where $P(B_m/A)$ *means,"probability that event occurs due to* B_m, *when it is known that A has occurred.*

Proof. Let $B_1, B_2, ..., B_n$ by any n mutually exclusive and exhaustive events and A is any other event on a random experiment.

$$\therefore A = (A \cap B_1) \cup (A \cap B_2) \cup ... \cup (A \cap B_n)$$

and

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + ... + P(A \cap B_n) \\ &= \sum_{m=1}^{n} P(A \cap B_m) \end{aligned}$$

Again from multiplication theorem

$$P(A \cap B_m) = P(A)P(B_m/A) \tag{10.1}$$

$$P(A \cap B_m) = P(B_m)P(A/B_m) \tag{10.2}$$

From(10.1)and (10.2),we get

$$\begin{aligned} P(A)P(B_m/A) &= P(B_m)P(A/B_m) \\ \therefore P(B_m/A) &= \frac{P(B_m)P(A/B_m)}{P(A)} \\ &= \frac{P(B_m)P(A/B_m)}{\sum_{m=1}^{n} P(A \cap B_m)} \end{aligned}$$

$$\therefore P(B_m/A) = \frac{P(B_m)P(A/B_m)}{\sum_{m=1}^{n} P(B_m)P(A/B_m)}$$

□

■ EXAMPLE 10.27

A bag contains 4 black and 1 white ball Another bag contains 5 black and 4 white balls. One ball has been taken out from one of the bags and found black. what is the probability that it came from first bags?

Solution: Let B_1=To select first bag, B_2=To select second bag, A = To select a ball, then

$$P(B_1) = \frac{1}{2}, P(B_2) = \frac{1}{2}$$

Probability of taking a black ball from first bag

$$P(A/B_1) = \frac{4}{5}$$

Probability of taking a black ball from second bag

$$P(A/B_2) = \frac{5}{9}$$

$\therefore$Required probability

$$\begin{aligned} p(B_1/A) &= \frac{P(B_1)P(A/B_1)}{P(B_1/A)P(A/B_1) + P(B_2)P(A/B_2)} \\ &= \frac{\frac{1}{2} \times \frac{4}{5}}{\frac{1}{2} \times \frac{4}{5} + \frac{1}{2} \times \frac{5}{9}} \\ &= \frac{36}{61} \end{aligned}$$

□

■ EXAMPLE 10.28

A purse contains there one rupees coins and four 50 paise coins. another purse contains four one rupee coins and five 50 paise. a one rupee coin has been taken out from one of the purses.Find the probability that it came from the first purse.

Solution: Let B_1=To taken a coin from first urn, B_2= To taken a coin from second urn and R =To take a rupee coin, then

$$P(B_1) = P(B_2) = \frac{1}{2}$$

Probability of taking a rupee coin from first urn,

$$P(R/B_1) = \frac{3}{7}$$

Probability of taking rupee coin from second urn,

$$P(R/B_2) = \frac{4}{9}$$

$\therefore$ Required Probability

$$\begin{aligned} P(B_1/R) &= \frac{P(B_1)P(R/B_1)}{P(B_1)P(R/B_1) + P(B_2)P(R/B_2)} \\ &= \frac{\frac{1}{2} \times \frac{3}{7}}{(\frac{1}{2} \times \frac{3}{7}) + (\frac{1}{2} \times \frac{4}{9})} \\ &= 0.49 \end{aligned}$$

□

Problems

10.19 There are 3 black and 4 white balls in a bag. Two balls are drawn one by one without replacement. If it is known that the second ball is then find the probability of the first ball being black.

10.20 An Insurance company insured 2,000 scooter drivers, 4,000 car drivers and 6,000 truck drivers. The probabilities to meat with an accident by a scooter driver, a car driver and a truck driver are 0.01,0.03 and 0.15 respectively .one of the insured drivers meats.

10.21 90 percent students nof a class are well prepared and 10 percent unprepared for an examination in mathematics. If the probability of passing with preparedness is 0.85 and that of not passing with unpreparedness. is 0.06 find the probability that a student selectd at random :

(i)who is found to have passed must have been well prepared, and

(ii)who is found to have faild must have been unprepared.

10.22 In a factory manufacturing fountain pens, machine A, B abd C manufacture 30%, 30% and 40% of the total production of fountain pens respectively of thier output 4%, 5% and 10% of the fountain pens are defective, if one fountain pen is selected at random and if it is found to be defective. What is the probability that it is manufactured by machine C?

UNIT III

UNIT 3
STATISTICAL TECHNIQUES - II

CHAPTER 11

PROBABILITY DISTRIBUTIONS

A listing of the probabilities for every possible value of a random variable is called a *probability distribution*. In this chapter, we discuss three very important distributions

1. Binomial Distribution
2. Poisson Distribution
3. Normal Distribution

11.1 BINOMIAL DISTRIBUTION

The binomial distribution[1] is a discrete probability distribution which is used when there are exactly two mutually exclusive outcomes of a trials. These outcomes are appropriately labeled *success* and *failure*. This is used to obtain the probability of observing x successes in n trials, with the probability of success on a single trial denoted by p. The binomial distribution assumes that p is fixed for all trials. Consider a random experiment which has two possible outcomes *success* and *failure*. The probability of success in each trial is p and remains same for all n trials. The trials are independent. Let X is a random variable which represents the number of successes in n trial. Then X has n values named as 0 success, 1 success, 2 successes, ..., n successes., i.e.

$$X = x : 0, 1, 2, 3, \ldots, n$$

[1]This distribution is associated with the name of James Bernoulli (1605-1705) and was published in 1713. It is also called as Bernoulli distribution in his honor.

with probability of x successes as

$$P(X=x) = {}^nC_x p^x q^{n-x}$$

where $q = 1-p$: the probability of failure in a single trial.

If the random experiment of n independent trials is repeated N time, then the *expected number* of trials showing a success is

$$N.{}^nC_x p^x q^{n-x}$$

The set of the numbers of success x along with the corresponding probability is known as *Binomial Probability Distribution*. The set of the number of successes along with the corresponding theoretical frequencies is called *Theoretical Binomial Distribution* with parameters n and p.

11.1.1 Assumptions of Binomial Distribution

Binomial distribution is based on the following assumptions

1. The Bernoulli trials are independent.
2. The number (n) of trial is finite and fixed.
3. The trials are repeated under identical conditions.
4. There are two mutually exclusive possible outcomes of each trial which are referred to as 'success' and 'failure'.
5. In a single trial probability of success is p and the probability of failure is q where $p+q=1$. The probability of success (and therefore for failure) remains same for each trial.

■ **EXAMPLE 11.1**

If on an average one ship in every ten is wrecked, find the probability of arrival of at least 4 ships safely out of 5 ship expected to arrive.

Solution: Given that one ship out of 10 is wrecked. i.e., 9 are safe.

$$\text{Probability of safe ship arriving:} = \frac{9}{10}$$

Here, we have to compute probability of arriving at least 4 ships safely out of 5 expected to arrive. Here n is 5. Since we need at least 4, it means there are either 4 safe ship or more than that. Thus here x has two values 4 and 5.

$$\begin{aligned} P(4 or 5) &= P(4)+P(5) \\ &= {}^5C_4 p^4 q^{5-4} + {}^5C_5 p^5 q^{5-5} \\ &= 5\left(\frac{9}{10}\right)^4\left(\frac{1}{10}\right) + \left(\frac{9}{10}\right)^5 = \frac{7}{5}\left(\frac{9}{10}\right)^4 \end{aligned}$$

□

■ **EXAMPLE 11.2**

10% of screws produced in a certain factory turn out to be defective. Find the probability that in a sample of 10 screws chosen at random, exectly two will be defective.

Solution: Here

$$p = 10\% = 0.1, \quad q = 1-p = 0.9, \quad n = 10, \quad x = 2$$

$$\begin{aligned} P(2) &=^n C_2 p^2 q^{n-2} \\ &=^1 0C_2 (0.1)^2 (0.9)^8 \\ &= 0.1937 \end{aligned}$$

□

■ EXAMPLE 11.3

Find out the Binomial Distribution to be expected by tossing 4 coins 320 times.

Solution: Let
X = Number of successes (i.e. = The number of heads when 4 coins are tossed),
n = 4, Number of coins,
p = Probability of getting a head on a coin = $\frac{1}{2}$ and,
$q = 1 - p = \frac{1}{2}$.
Possible values of X are : $0, 1, 2, 3, 4$. The corresponding probabilities are obtained in the expansion of $(p+q)^n$:

$$(p+q)^4 = p^4 + ^4C_1 p^3 q + ^4C_2 p^2 q^2 + ^4C_3 p q^3 + q^4$$

$$\left(\frac{1}{2}+\frac{1}{2}\right)^4 = \left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) + 6\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 + 4\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4$$
$$= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16}$$

Thus, we have

$X = x$	0	1	2	3	4
$P(X = x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Expected Frequency = $NP(X = x)$,where N =320. The required theoretical Binomial Distribution is as follows:

$X = x$	0	1	2	3	4
Expected frequency	20	80	120	80	20

□

11.1.2 Mean, Variance and Standard deviation of Binomial distribution

Let the Binomial distribution be

$$p(x) = P(X = x) =^n C_x p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

$$\begin{aligned}
\text{Mean, } \bar{X} = E(X) &= \sum_{x=0}^{n} x p(x) \\
&= \sum_{x=0}^{n} x^n C_x p^x q^{n-x} \\
&= 0.q^n + 1.^nC_1 q^{n-1} p + 2.^nC_2 q^{n-2} p^2 + \ldots + np^n \\
&= 0 + nq^{n-1} p + \frac{n(n-1)}{1!} q^{n-1} p^2 + \ldots + np^n \\
&= np[q^{n-1} + (n-1)q^{n-2} p + \ldots + p^{n-1}] \\
&= np(q+p)^{n-1} \\
&= np\{\because (q+p) = 1\}
\end{aligned}$$

Variance,

$$\begin{aligned}
\sigma^2 &= E[X - E(X)]^2 \\
&= E(X^2) - (E(X))^2
\end{aligned}$$

Now

$$\begin{aligned}
E(X^2) &= \sum_{x=0}^{n} x^2 p(x) \\
&= \sum_{x=0}^{n} x^2 ({}^nC_x p^x q^{n-x}) \\
&= \sum_{x=0}^{n} [x(x-1) + x].^nC_x p^x q^{n-x} \\
&= \sum_{x=0}^{n} x(x-1).^nC_x p^x q^{n-x} + \sum_{x=0}^{n} x.^nC_x p^x q^{n-x} \\
&= \sum_{x=2}^{n} n(n-1)p^2.^{n-2}C_{x-2} p^{x-2} q^{n-x} + np\} \\
&= n(n-1)p^2 (q+p)^{n-1} + np \\
&= n(n-1)p^2 + np\{\because q+p = 1\} \\
\sigma^2 &= E(X^2) - (E(X))^2 \\
&= n(n-1)p^2 + np - (np)^2 \\
&= n^2p^2 - np^2 + np - n^2p^2 \\
&= np - np^2 = np(1-p) = npq
\end{aligned}$$

$\therefore$ Standard Deviation

$$\sigma = \sqrt{\text{Variance}} = \sqrt{npq}$$

11.1.3 Constants of Binomial Distribution

Let the Binomial Distribution be

$$P(X = x) = nC_x p^x q^{n-x}, \;\; x = 0, 1, 2, \ldots, n$$

Mean,

$$\mu = E(X) = np$$

Standard deviation,

$$\sigma = \sqrt{npq}$$

First central moment,

$$\mu = 0$$

Second central moment,

$$\mu_2 = \sigma^2 = npq$$

Third central moment,

$$\mu_3 = npq(q-p)$$

Fourth central moment,

$$\mu_4 = 3n^2p^2q^2 + npq(1-6pq)$$

Karl Pearson coefficients,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(q-p)^2}{npq}, \gamma_1 = +-\sqrt{\beta}$$

The sign of γ_1 depends on μ_3. If $\mu_3 > 0$ then $\gamma_1 = +\sqrt{\beta 1}$. If $\mu_3 < 0$ then $\gamma_1 = -\sqrt{\beta_1}$

$$\beta_2 = \frac{\mu_4}{\mu_4^2} = 3 + \frac{1-6pq}{npq}$$

$$\gamma_2 = \beta_2 - 3 = \frac{1-6pq}{npq}$$

11.1.4 Skewness:

- If $\mu_3 = npq(q-p) < 0$,negative i.e, p=q then no skewness is negative.
- If $\mu_3 = npq(q-p) = 0$,i.e,p=q then no skewness the distribution is symmetrical.
- If $\mu = npq(q-p) > 0$,positive i,e. p¡q then skewness is positive.

11.1.5 Kurtosis

- If $\beta_2 = 3 + \frac{1-6pq}{npq} < 3$i,e.,$1-6pq < 0$ i,e.,$pq > \frac{1}{6}$,the distribution is platykurtic.
- If $\beta_2 = 3 + \frac{1-6pq}{npq} = 3$i,e.,$pq = \frac{1}{6}$,the distribution is mesokurtic.
- If $\beta_2 = 3 + \frac{1-6pq}{npq} > 3$ i,e.,$pq < \frac{1}{6}$,the distribution is leptokurtic.

11.1.6 Mode

The value of x for which $P(X = x)$ is maximum is called mode.

Case 1 If np = integer,then mode = np = mean

Case 2 If $np + p$ = integer, then there are two modes $np + p$ and $np + p - 1$. where $p(np+p) = p(np+p-1)$

Case 3 .If $np + p = k + d$ where k is an integer and d is a fractional,then mode = k.

■ EXAMPLE 11.4

The mean of a binomial distribution is 4 and third central moment is 0.48. Find standard deviation and mode of the distribution.

Solution: Given:

$$\text{Mean } = np = 4 \tag{11.1}$$

$$\mu_3 = npq(q-p) = 0.48 \tag{11.2}$$

Dividing equation (11.2) by (11.1), we get

$$\begin{aligned}
\frac{npq(q-p)}{np} &= \frac{0.48}{4} = \frac{48}{400} = \frac{12}{100} = \frac{3}{25} q(q-p) \\
&= \frac{3}{25} \\
q[q-(1-q)] &= \frac{3}{25} \\
q(q+q-1) &= \frac{3}{25} \\
2q^2 - q &= \frac{3}{25} \\
50q^2 - 25q - 3 &= 0 \\
50q^2 - 30q + 5q - 3 &= 0 \\
10q(5q-3) + 1(5q-3) &= 0 \\
(5q-3)(10q+1) &= 0 \\
5q - 3 = 0 &\text{ or } 10q + 1 = 0 \\
q = \frac{3}{5} &\text{ or } q = -\frac{1}{10} \\
q = \frac{3}{5} &\{\text{since q can not be negative}\}
\end{aligned}$$

Hence

$$p = 1 - q = 1 - \frac{3}{5} = \frac{2}{5}$$

and $n = \frac{4}{q} = 4 \times \frac{5}{2} = 10$

Now Standard Deviation

$$= \sqrt{npq} = \sqrt{10 \times \frac{3}{5} \times \frac{2}{5}} = \sqrt{2.4} = 1.55$$

$$np + p = 10 \times \frac{2}{5} + \frac{2}{5} = 4 + \frac{2}{5}$$

$$\text{Hence Mode } = 4 \qquad \{\text{Integral part of } (np+p)\}$$

□

11.1.7 Characteristics of Binomial Distribution

1. Binomial distribution is a discrete probability distribution based on Binomial Theorem.

2. Theoretical Binomial distribution can be obtained by multiplying the probabilities by N.

3. The shape and location of binomial distribution changes as p the probability of success changes for a given n. When $p = q = 1 - p$ the distribution is symmetrical. When the number of success are in ascending order, the distribution is negatively skewed for $p > 0.5$ (or $p > q$) and the distribution is positively skewed for $p < 0.5$(or $p < q$). That is binomial distribution is asymmetrical for $p \pm q$.

   This distribution is platykurtic for $pq > \frac{1}{6}$, mesokurtic for $pq = \frac{1}{6}$ and lepokurtic for $pq = \frac{1}{6}$.

4. If n is large and if neither p nor q is too close zero i,e., the difference between p and q is very small, then the binomial distribution tends to symmetry and may be approximately *normal distribution*[2].
5. The constants of binomial distribution $P(X=x) = ^nC_x p^x q^{n-x}, x = 0, 1, 2, \ldots$ are as follows: Mean = np, Standard deviation = $\sqrt{npq}$, First central moment, $\mu, = 0$, Second central moment, $\mu_2 = npq = Variance$, Third central moment, $\mu_3 = npq(q-p)$, Fourth central moment, $\mu_4 = 3n^2p^2q^2 + npq(1-6pq)$, $\beta_1 = \frac{(q-p)^n}{npq}$, $\beta_2 = 3 + \frac{(1-6pq)}{npq}$.

11.1.8 Application of Binomial Distribution

The use of binomial distribution is made in case of division by dichotomy where one event is called success and another event is called failure. Under certain assumptions and circumstances the nature of binomial distribution is known. It is applicable to the situations of random sampling from finite population with replacement or sampling from an infinite population with or without replacement. The binomial distribution is very useful in decision making situations in business. One area in which it has been very widely applied is quality control. Binomial distribution is considered an important tool in forecasting of the events based on random sampling.

Problems

11.1 The incidence of occupational disease in an industry is such that the workman have a 20% chances of suffering from it. What is the probability that out of six workman 4 or more will contact disease?

11.2 The normal rate of infection of a certain disease in animal is known to be 40%. In an experiment with 6 animals caught infection. Find the probability of the observed result.

11.3 In a binomial distribution consisting of 5 independent trials, probabilities of 1 and 2 success are 0.4096 and 0.2048 respectively. Find the parameter p of this binomial distribution.

11.4 Six dice are thrown 729 times. How many times do you expect at least three dice to show a five or six?

11.5 Eight coins are tossed and this experiment is repeated 100 times. Find binomial distribution and calculate its means and standard deviation.

11.6 Assuming that 50% of a population in a town reads newspapers and further assuming that 1024 investigators each take 10 individuals to find out if they read newspapers,how many investigators would you expect to report that three or less people read newspapers.

11.7 Assuming that half the population is vegetarian so that chance of an individual being a vegetarian is $\frac{1}{2}$ and assuming that out of 100 investigators each takes a sample of 10 individuals to see whether they are vegetarian. How many investigators would expect to report that three or less people are vegetarian?

11.8 In an army battalion $3/5$ of the soldiers are married and the remainder $2/5$ unmarried. What is the probability of being $0, 1, 2, 3, 4, 5$ married soldiers in a row of 5 soldiers?

11.9 Find all the expected frequencies of a theoretical binomial distribution $128(\frac{1}{2}+\frac{1}{2})^5$

11.10 Four dice were thrown 64 times. Spot 3 or 4 was considered a success and spot 1 or 2 a failure. The outcome of a 5 or 6, was not considered a trial. The results were:

[2]Discussed in next sections.

Number of success	0	1	2	3	4
Frequency	0	5	13	22	24

11.11 Find the mean and standard deviation of the observed frequency distribution and the theoretical binomial distribution.

11.12 For Theoretical Binomial distribution $\bar{x} = np = 4 * \frac{1}{2} = 2$

11.13 Assuming head as a success in 512 tosses of 8 coins together find the expected frequencies obtaining the probabilities of various successes.

No. of Success x	Probability p(x)	Expected Frequency Np(x)
0	1/256	2
1	8/256	16
2	28/256	56
3	56/256	112
4	70/256	140
5	56/256	112
6	28/256	56
7	8/256	16
8	1/256	2
Total	1	512

11.14 A set of 8 coins is thrown 256 times and the number of heads appearing in each throw is recorded. The results obtained are given in the following table. Estimate the mean number of heads and then calculate the expected frequencies using this estimate if the binomial law holds :

No.of heads	0	1	2	3	4	5	6	7	8
No. of throws	2	6	38	52	59	56	32	10	1

11.15 There are 64 beds in a garden and 3 seeds of particular type of flower are sown in each bed. The probability of a flower being blue is 1/4. Find the number of beds with 3,2,1 and 0 blue flowers.

11.16 In 104 liters of 4 mice, the number of litters which contain 0,1,2,3,4 female mice is as follows:

No. of female mice,x	0	1	2	3	4
No. of litters,f	8	28	34	24	10

Estimate the probability of a mouse being a female and find the expected frequencies.

11.17 Obtain the mean and mode for the following binomial distribution if

a) n=99, p=0.6
b) n=60, p=1/6
c) n=8 , p=1/5

11.18 In a binomial distribution mean is 3 and variance is 2, find the remaining constants ($\mu_1, \mu_2, \mu_3, \mu_4, \beta_1, \beta_2$).

11.19 4 coins are tossed 160 times and following observed frequency distribution:

No. of heads	0	1	2	3	4
No. of tosses	17	32	54	51	6

Assuming that the coins are unbiased calculate the expected frequencies for the number of heads and test the goodness of fit.

11.20 A set of 5 coins was tossed 96 times. The following table gives the observed frequency distribution of the number of heads in a single trial. Calculate the expected frequencies on the assumption that the coins were unbiased. use — test to examine the correctness of this assumption (x=number of coins that show head in a single trial, f=No. of throws)

x	5	4	3	2	1	0
f	7	19	35	24	8	3

11.21 The observed frequency distribution of 128 throws of 7 coins, according to the number of heads. Fit a binomial distribution under the assumption that the coins are unbiased. What are the mean and standard deviation of the fitted distribution?

No. of heads	0	1	2	3	4	5	6	7
No of throws	7	6	19	35	30	23	7	1

11.22 Criticize the statement: The mean of a binomial distribution 8 and standard deviation is 3.

11.2 POISSON DISTRIBUTION

Let X be a discrete random variable where probability of $X = x$ is given by

$$P(X=x) = \frac{e^{-m}m^x}{x!},\ x = 0,1,2......., \quad m \geq 0$$

the probability distribution of X is said to be Poisson distribution[3] with parameter m. Here m is mean of the distribution. Poisson distribution is a limiting form of binomial distribution where p is very small and n is large enough such that $np = m$ (Constant).

Theorem 13. *In binomial distribution $(p+q)^n$, when p is very small, i.e., $p \to 0$ and n is very large, i.e., $n \to \infty$, in such a way that $np = m$ is some fixed number, then the probability of x successes is*

$$P(X=x) = \frac{e^{-m}m^x}{x!}$$

Proof. Probability of x successes in binomial distribution is

$$\begin{aligned} P(X=x) &= {}^nC_x q^{n-x}p^x \\ &= \frac{n!}{x!(n-x)!}(1-p)^{n-x}p^x \\ &= \frac{n!}{x!(n-x)!}\left(1-\frac{m}{n}\right)^{n-x}\left(\frac{m}{n}\right) \\ &= \frac{m^x}{x!}\left(1-\frac{m}{n}\right)^n \frac{n!}{(n-x)!\left(1-\frac{m}{n}\right)^x} \end{aligned}$$

when $n \to \infty$

$$= \frac{m^x}{x!}e^{-m}$$

[3]This distribution was developed by S.D. Poisson in 1837.

as when n tends to infinity, we have

$$\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n = e^{-m}$$

$\square$

When the expected frequencies for an observed frequency distribution are obtained using Poisson Law, then general expected frequency is given by $N.e^{-m}\frac{m^x}{x!}$, where N= total observed frequencies. The theoretical distribution so obtained is called *theoretical Poisson distribution*.

Poisson distribution is applicable in case of a number of random situations with a small probability of success. Poisson distribution is used in place of binomial distribution when p is very small,n is very large and $np = m$ is fixed number.

Some very common cases where Poisson distribution is used: The number of deaths at a place in any year by an epidemic or by a rare disease;

1. The number of mistakes per page by an experienced typist;
2. The number of typographical errors per page in a book printed in a good press;
3. The number of defective items in the items manufactured by a big factory;
4. The number of bacteria in a drop of clean water;
5. The number of suicides per day in a city;
6. The number of telephone calls per minute in an telephone booth.

11.2.1 Mean, Variance and Standard deviation of Poisson distribution

Mean:

$$\begin{aligned}\bar{X} &= \frac{\sum_{x=0}^{\infty} x.p(x)}{\sum_{x=0}^{\infty} p(x)} = \sum_{x=0}^{\infty} xp(x)\\ &= \sum_{x=0} x.e^{-m}\frac{m^x}{x!}\\ &= e^{-m}.m\sum_{x=1}\frac{m^{x-1}}{(x-1)!}\\ &= e^{-m}.m.e^m = m\end{aligned}$$

Second Moment about origin :

$$\begin{aligned}\mu_2' &= \sum_{x=0}^{\infty} x^2 p(x)\\ &= \sum_{x=0}^{\infty}[x(x-1)+x]e^{-m}\frac{m^x}{x!}\\ &= \sum_{x=0}^{\infty} x(x-1)e^{-m}\frac{m^x}{x!} + \sum_{x=0}^{\infty} xe^{-m}\frac{m^x}{x!}\\ &= m^2e^{-m}\sum_{x=2}\frac{m^{x-2}}{(x-2)!} + \bar{x}\\ &= m^2e^{-m}.e^m + m\\ &= m^2 + m\end{aligned}$$

Second Moment about mean:

$$\mu_2 = \mu_2' - (\mu_1')^2$$
$$= m^2 + m - m^2$$
$$= m$$

Variance:

$$\sigma^2 = \mu_2 = m$$

Standard Deviation:

$$\sigma = \sqrt{Variance} = \sqrt{m}$$

11.2.2 Constants of Poisson Distribution

Mode:
case 1: If m is a positive integer, then there are two modes at $x = x - 1$ and $x = m$.
case 2: If m is not a positive integer, then the integral part of m is mode.

Third moment about mean, $\mu_3 = m$, always positive

Fourth moment about mean, $\mu_4 = 3m^2 + m$ or $m(3m+1)$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m}, \gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}}$$
$$\beta_1 = \frac{\mu_4}{\mu_2^2} = \frac{m(3m+1)}{m^2} = \frac{3m+1}{m} = 3 + \frac{1}{m}$$
$$\gamma_2 = \beta_2 - 3 = 3 + \frac{1}{m} - 3 = \frac{1}{m}$$

Remark: The advantage of Poisson distribution is that if we know the value of the parameter m, we can find all the other constants of this distribution.

■ **EXAMPLE 11.5**

The parameter m of a Poisson distribution is 3.24. Find the constants of this distribution.

Solution:

$$P(X = x) = p(x) = e^{-m}\frac{m^x}{x!}, x = 0, 1, 2, ...$$
$$p(x) = e^{-3.24}\frac{(3.24)^x}{x!}, x = 0, 1, 2, ...$$

Mean ,

$$\bar{X} = m = 3.24$$

Variance,

$$\sigma^2 = m = 3.24 =$$

Second central moment,

$$\mu_2$$

Standard Deviation ,

$$\sigma = \sqrt{m} = \sqrt{3.24} = 1.8$$

Third central moment,

$$\mu_4 = (3m+1) = 3.24[3(3.24)+1] = 3.24(9.72+1) = 3.24 \times 10.72 = 34.7328$$

Moment coefficient of Skewness:.

$$\beta_1 = \frac{1}{m} = \frac{1}{3.24} = .3086$$

$$\gamma_1 = \sqrt{\frac{1}{m}} = \sqrt{\frac{1}{3.24}} = \sqrt{0.3086} = 0.5556$$

Moment coefficients of Kurtosis:.

$$\beta_2 = 3 + \frac{1}{m} = 3 + \frac{1}{3.24} = 3 + .3086 = 3.3086$$

$$\gamma_2 = \beta_2 - 3 = \frac{1}{m} = \frac{1}{3.24} = .3086$$

Mode:

$$m = 3.24 = 3 + 0.24$$

Since m is not a positive integer, hence mode = 3 □

■ **EXAMPLE 11.6**

In a certain Poisson frequency distribution the frequency corresponding to 2 successes is half the frequency corresponding to 3 successes. Find its mean and standard deviation.

Solution: Let the total frequency be N. Then the frequency for x successes in the Poisson distribution is

$$N.e^{-m}.\frac{m^x}{x!}, \quad x = 0, 1, 2, 3.........$$

Frequency of 2 successes $= N.e^{-m}\frac{m^2}{2!}$

Frequency of 3 successes $= N.e^{-m}\frac{m^3}{3!}$

According to the condition given:

$$Ne^{-m}.\frac{m^2}{2!} = \frac{1}{2}[Ne^{-m}\frac{m^3}{3!}]$$

$$\frac{1}{2\times 1} = \frac{1}{2} \times \frac{m}{3\times 2\times 1}$$

$$m = \frac{2\times 3\times 2\times 1}{2\times 1} = 6$$

$\therefore$ Mean, $\bar{X} = m = 6$ and Standard deviation $\sigma = \sqrt{m} = \sqrt{6} = 2.449.$ □

11.2.3 Computation of Probabilities and Expected Frequencies in Poisson distribution

In a Poisson distribution the probability of x successes (or probability mass function of X) is given by

$$p(x) = e^{-m}.\frac{m^x}{x!}, x = 0, 1, 2, ...$$

To find the probabilities for $x = 0, 1, 2, ...$ first of all we obtain the value of e^m.The values of e^{-m} for the various values of m are given in a table from which they can be read easily or the values of m are given in a table from which they can be read easily or the values may be computed as follows:

$$e^{-2} = \text{Antilog}[-2log_{10}e] = \text{Antilog}[-2log_{10}2.718] = 0.1353$$

After finding the value of e^{-m} or p(0) we find p(1), p(2), ... etc. By the relation

$$p(x) = e^{-m}.\frac{m^x}{x!}$$

or by the relation

$$p(x+1) = \frac{m}{x+1}p(x)$$

Expected frequencies are obtained as follows:
Expected frequency of x successes = $Np(x)$.

m	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	.9048	.8958	.8869	.8681	.8694	.8607	.8521	.8437	.8353	.8207
0.2	.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	.5488	.5434	.5379	.5326	.5273	.5220	.5169	.4117	.5066	.5016
0.7	.4966	.4916	.4868	.4819	.1771	.4724	.4677	.4630	.4548	.4538
0.8	.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	.4066	.4025	.3985	.3942	.3906	.3867	.3829	.3791	.3753	.3716

The value of e^{-m} for m = 1 to m = 10

m	1	2	3	4	5
e^{-m}	0.36788	0.13534	0.04979	0.01832	0.006738
m	6	7	8	9	10
e^{-m}	0.002479	0.000912	0.000335	0.000123	0.000045

To read the value of e^{-m} for $m = 0.51$ from the above table, we take 0.5 in the column (m) and then read the value in column (1) in front of 0.5. Thus

$$e^{-0.51} = .6005$$

Similarly $e^{-0.61} = .5434$ and $e^{-2.51} = e^{-2.0-0.51} = .13534 \times .6005 = .08127$

■ EXAMPLE 11.7

Suppose that a manufactured product has 2 defects per unit if product is inspected. Using Poisson distribution calculate the probabilities of finding a product (i) without any defect , (ii) 3 defects (iii) with 4 defects , (iv) with 3 defects or 4 defects . (Given $e^{-2} = 0.135$).

Solution: Let X = Number of defects per unit, and

$$p(x) = P(X = x) = e^{-m}\frac{m^x}{x!},\ \ x = 0, 1, ...$$

Here $m = 2$

$$p(0) = e^{-2} = 0.135 \ \ (given)$$

$$p(1) = p(0) \times m = 0.135 \times 2 = 0.27$$

$$p(2) = p(1) \times \frac{m}{2} = 0.27 \times \frac{2}{2} = 0.27$$

$$p(3) = p(2) \times \frac{m}{3} = 0.27 \times \frac{2}{3} = 0.18$$

$$p(4) = p(3) \times \frac{m}{4} = 0.18 \times \frac{2}{4} = 0.09$$

In this way
(i) The probability of no defect = p(0) = 0.135
(ii) The probability of 3 defects = p(3) = 0.18
(iii) The probability of 4 defects = p(4) = 0.09
(iv) The probability of 3 or 4 defects = 0.18+0.09=0.27 □

■ EXAMPLE 11.8

2 percent of the electric bulbs produced by a company are defective.Find the probability that in a sample of 200 bulbs (i)less than bulbs are defective and (ii) more than 3 bulbs are defective. (Given $e^{-4} = 0.0.183$)

Solution:

$$n = 200, p = 2 = \frac{2}{100} = .02$$
$$m = np = 200 \times .02 = 4$$
$$P(X = x) = e^{-m}\frac{m^x}{x!}, x = 0, 1, 2, \ldots\ldots..$$
$$P(X = x) = e^{-4}\frac{4^x}{x!}, x = 0, 1, 2, \ldots\ldots\ldots.$$

(i) P (less than 2 bulbs are defective)

$$= P(X < 2) = P(X = 0) + P(X = 1) = e^{-4} + e^{-4}.\frac{4}{1!}$$
$$= e^{-4}(1+4) = 5e^{-4} = 5 \times 0.0183 = 0.0915$$

(ii) P (more than 3 bulbs are defective)

$$= P(X > 3) = 1 - P(X < 3) = 1 - [e^{-4} + e^{-4}.\frac{4}{1!} + e^{-4}.\frac{4^2}{2!} + e^{-4}.\frac{4^3}{3!}]$$
$$= 1 - e^{-4}[1 + 4 + 8 + \frac{32}{3}] = 1 - .0183 \times \frac{71}{3} = 1 - .4331 = .5669$$

□

■ EXAMPLE 11.9

The distribution of accidents by taxi drivers in city in a year is a Poisson Distribution. The mean of the distribution is 3. Find out the member of taxi drivers out of 1000 taxi drivers (i) who are not involved in any accident in a year and (ii) who are involved in more than 3 accidents in a year.

Solution: X = Number of accidents in a year m = 3, the mean of Poisson Distribution N = 1000 , number of taxi driver $p(x) = P(X = x) = e^{-m}\frac{m^x}{x!}, x = 0, 1, 2, 3, \ldots\ldots$

X=x	Probability, P(X=x),p(x)	Expected Number , Np(x)
0	$e^{-m} = e^{-3} = 0.0498$	1000×0.0498=49.8
1	$me^{-m} = 3e^{-3} = 0.1494$	1000×0.1494=149.4
2	$\frac{m^2e^{-m}}{2!} = \frac{9}{2}e^{-3} = 0.2241$	1000×0.2241=224.1
3	$\frac{m^3e^{-m}}{3!} = \frac{27}{6}e^{-3} = 0.2241$	1000×0.2241=224.1

Figure

(i) $Np(0) = 49.8 - 50 \approx 0$, no accident
(ii) $N - N[p(0) + p(1) + p(2) + p(3)]$, more than 3 accidents

$$= 1000 - (49.8 + 149.4 + 224.1 + 224.1)$$
$$= 1000 - 647.4 = 352.6 - 35.3$$

□

■ EXAMPLE 11.10

A manufacturer of pins knows that on an average 5% of his product is defective.He sells pins in boxes of 100 and guarantees that not more than 4 pins will be defective. What is the probability that a box will meet the guaranteed quality?($e^{-5} = .0067$)

Solution: The probability that pin is defective , p = 5 % or 0.05
Let X = The number of defective pins in a packets of 100 pins, $p = 0.05$ and $n = 100$,

$$m = n \times p = 0.05 \times 100 = 5$$

and

$$p(x) = P(X = x) = e^{-m}.\frac{m^x}{x!}$$

$$\begin{aligned}
\therefore p(0) &= e^{-5} = .0067\\
p(1) &= p(0) \times \frac{m}{1} = .0067 \times \frac{5}{1} = .0335\\
p(2) &= p(1) \times \frac{m}{2} = .0335 \times \frac{5}{2} = .0838\\
p(3) &= p(2) \times \frac{m}{3} = .0838 \times \frac{5}{3} = .1397\\
p(4) &= p(3) \times \frac{m}{4} = .1397 \times \frac{5}{4} = .1746
\end{aligned}$$

P(a box will meet the guaranteed quality)

$$\begin{aligned}
&= P \text{ (The number of defective pins is not more than 4)}\\
&= 1 - p \text{ (The number of defective pins is 4 or less)}\\
&= 1 - [p(0) + p(1) + p(2) + p(3) + p(4)]\\
&= 1 - [0.0067 + 0.0335 + 0.0838 + 0.1397 + 0.1746]\\
&= 0.5617
\end{aligned}$$

□

■ EXAMPLE 11.11

In a certain factory turning out razor blades, there is a small chance of 1/500 for any blade to be defective. The blades are supplied in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing (1) no defective, (2)one defective and (3) two defective blades, respectively in a consignment of 10000packets.

Solution:

Here,

$$n = 10,\ p = \frac{1}{500},\ N = 10000$$

$$\therefore m = np = 10 \times \frac{1}{500} = .02$$

Let X = The number of defective blades in a packet $(= 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$.

According to the Poisson distribution,

$$p(x) = P(X = x) = e^{-m}\frac{m^x}{x!},\ x = 0, 1, 2, \ldots\ldots\ldots$$

$$P(0) = e^{-0.02} = .9802$$

$$\begin{aligned}
p(x+1) &= p(x) \times \frac{m}{x+1} \text{ gives,} \\
p(1) &= p(0) \times \frac{0.02}{1} = 0.9802 \times \frac{0.02}{1} = 0.019604 \\
p(2) &= p(1) \times \frac{0.02}{2} = 0.019604 \times \frac{0.02}{2} = 0.00019604 \\
p(3) &= p(2) \times \frac{0.02}{3} = 0.00019604 \times \frac{0.02}{3} \approx 0
\end{aligned}$$

Required expected frequencies:

(1) No defective $e_0 = Np(0) = 10000 \times 0.9802 = 9802.0 = 9802$

(2) One defective $e_1 = Np(0) = 10000 \times 0.019604 = 196.04 = 196$

(3) Two defective $e_2 = Np(2) = 10000 \times 0.00019604 = 1.9604 = 2$ □

■ EXAMPLE 11.12

A car hire firm has two cars which it hires out day by day. The number of demands for a car on each day is distributed as Poisson distribution with mean 1.5. Calculate the proportion of days on which (i) neither car is used, and (ii) some demand is refused.

Solution:

Let X = demand of a car, then

$$p(X = x) = e^{-m}\frac{m^x}{x!},\ \text{where } m = 1.5.$$

(i) Neither car is used i.e., the demand is zero, required probability (or proportion)

$$p(X = 0) = e^{-m} = e^{-1.5} = .2231$$

(ii) The demand is refused, means the demand is more than 2: The required probability (or proportion)

$$\begin{aligned}
p(X > 2) &= 1 - p(X < 2) \\
&= 1 - [p(0) + p(1) + p(2)] \\
&= 1 - \{e^{-m} + e^{m}m + e^{-m}\frac{m^2}{2!}\} \\
&= 1 - \left[0.2231 + 0.2231 \times 1.5 + (0.2231) \times \frac{(1.5)^2}{2}\right] \\
&= 0.1913
\end{aligned}$$

□

EXAMPLE 11.13

Mean of a distribution which is explained by the Poisson distribution is 2.0. Find out the probabilities corresponding to various successes 0,1,2,3........ If the total frequency is 100 what are the various frequencies?

Solution:

Probability of r successes

$$p(r) = e^{-m}\frac{m^r}{r!}, \text{ where } m = 2$$

$$\begin{aligned}\therefore p(0) &= e^{-m}\frac{m^0}{0!} = e^{-m} = e^{-2}\\ &= 2.71828^{-2} = 0.1353\end{aligned}$$

From formula $p(r+1) = p(r)\times\frac{m}{r+1}$, we get,

$$\begin{aligned}p(1) = p(0)\times\frac{m}{1} &= 2p(0) = 2\times .1353 = .2706\\ p(2) = p(1)\times\frac{m}{2} &= \frac{2}{2}\times .2706 = .2706\\ p(3) = p(2)\times\frac{m}{3} &= \frac{2}{3}\times .2706 = .1804\\ p(4) = p(3)\times\frac{m}{4} &= \frac{2}{4}\times .1804 = .0902\\ p(5) = p(4)\times\frac{m}{5} &= \frac{2}{5}\times .0902 = .03608\\ p(6) = p(5)\times\frac{m}{6} &= \frac{2}{6}\times .03608 = .0120266\\ p(7) = p(6)\times\frac{m}{7} &= \frac{2}{7}\times .01203 = .0034371\end{aligned}$$

The computations may be carried out up to infinite successes, but it is clear that the probabilities are decreasing very fast and will become very low after some time, then the computation is meaning less. Thus the Poisson distribution is as follows:

x	0	1	2	3	4	5	6	7 or more
$p(x)$	. 1353	. 2706	. 2706	. 1804	. 0902	. 0361	. 0120	. 0034

We have not computed the probabilities after p(7), the last probability is for '7 or more than 7'. this is obtained by subtracting the sum $[p(0)+P(1)+p(2)+p(3)+p(4)+p(5)+p(6)]$ from 1.

If the total frequency is 100,then the expected frequencies are as follows :

x	$Np(x)$
0	100×.1353=13.53=14
1	100×.2706=27.06=27
2	100×.2706=27.06=27
3	100×.1804=18.04=18
4	100×.0902=9.02=9
5	100×.0361=3.61=4
6	100×.0120=1.20=1
7 or more	100×.0034=0.34=0

□

EXAMPLE 11.14

The following table give the data regarding the deaths of soldiers from kicks of horses per army corps per year for General Army Corps:

Deaths	0	1	2	3	4
Frequency	109	65	22	3	1

Fit a Poisson distribution and calculates expected frequencies.

Solution:

Death x	Frequency f	fx
0	109	0
1	65	65
2	22	44
3	3	9
4	1	4
Total	200	122

Mean ,

$$m = \frac{\sum fx}{\sum f} = \frac{122}{200} = 0.61, e^{-m} = e^{-0.61} = 0.5434$$

$$P(X = x) = e^{-m}\frac{m^x}{x!}, x = 0, 1, 2, \ldots.$$

Formula :

$$p(x+1) = p(x) \times \frac{m}{x+1}, gives$$

$$p(1) = p(0) \times \frac{0.61}{1} = 0.5434 \times 0.61 = 0.3314$$

$$p(1) = p(2) \times \frac{0.61}{2} = 0.3314 \times \frac{0.61}{2} = 0.1011$$

$$p(3) = p(2) \times \frac{0.61}{3} = 0.1011 \times \frac{0.61}{3} = 0.0205$$

$$p(4) = p(3) \times \frac{0.61}{4} = 0.0205 \times \frac{0.61}{4} = 0.0031$$

To make the total probability 1, we can find the probability of '4 and more' instead of probability of 4. p (4 or more)

$$= 1 - (p(0) + p(1) + p(2) + p(3))$$

$$= 1 - (.5434 + .3314 + .1011 + .0205)$$

$$= 1 - (.9964) = .0036$$

Remark: The number of deaths of German soldiers in 20 units of German Army Corps within 20 years from kicks of horses is an important example of Poisson distribution. Probabilities and Expected frequencies are as follows:

x	$p(x)$	$Np(x)$
0	0.5434	200 × 0.5434=108.7=109
1	0.3314	200 × 0.3314=66.3=66
2	0.1011	200 × 0.1011=20.2=20
3	0.0205	200 × 0.0205=4.1=4
4 or more	0.0031	20 × 0.0031=0.6=1
Total	0.9995=1	200

□

Problems

11.23 If the proportion of defective items in a bulk is 4 percent find the probability of not more than two defective in a sample of 10.(It is known that $e^{-4} = .6703$)

11.24 The past experience shows that there occur on an average 4 industrial accidents per month.Find the probability of the occurrence of less than 4 industrial accidents in a certain month. (Given : $e^{-4} = .018$)

11.25 In a town 10 accidents take place in a span of 50 days.Assuming that the number of accidents follow Poisson distribution , find the probability that there will be three or more accidents in a day.

11.26 The average number of customers per minute arriving at a shop is 5. Using Poisson distribution find the probability that during one particular minute exactly 6 customers will arrive .

11.27 After correcting the errors of the first 50 pages of a book, it is found that on the average there are 3 errors per 5 pages . Use Poisson distribution to estimate the number of pages with 0,1,2,3, errors in the whole book of 1,000 pages (Given$e^{-.6} = .5488$)

11.28 A producer claims that 5% of his product is defective. He sells his product in boxes of 100 and guarantees that not more than 3 items will be defective . What is the probability that a box will fail to meet the guarantee?

11.29 A telephone exchange receives on an average 4 calls per minute. Find the probability on the basis of Poisson distribution of receiving:

- **a)** 2 or less calls per minute,
- **b)** up to 4 calls per minute, and
- **c)** more than 4 calls per minute. (Given: $e^{-4} = .0813$)

11.30 The number of telephone call an operator receives from 8 to 5 minute past 8 follows a Poisson distribution with m = 3 . What is the probability that the operator (i) will not receive a telephone call and (ii) exactly 3 calls in the same time interval tomorrow?

11.31 The probability that a particular injection will have reaction to an individual is 0.002. Find the value probability that out of 1000 persons (×) exactly 2 individuals and (×) at least one individual will have a reaction from injection.

11.32 In a certain factory turning out optical lenses there is a small chance $\frac{1}{500}$for any lens to be defective. The lenses are supplied in packets of of 10. Use Poisson distribution to calculate the approximate number of packets containing (i) no defective, (ii) one defective, (iii) two defective (iv) three defective and (v) four defective lenses respectively in a consignment of 20000 packets. (given $e^{-.02} = .9802$)

11.33 In factory manufacturing fountain pen, the chance having a defective pen is 0.5 percent, 100 pens are kept in box. What is the percentage of boxes in which (i) there is no defective pen (ii) there is at least one defective pen and (iii)there are 2 or more than 2 defective pen?(given $e^{-.5} = .6065$)

11.34 A manager accepts the work submitted by his typist only when there is no mistake in the work. The typist has to type on an average 20 letters per day of about 200 words each. Use Poisson distribution to find the chance of his making a mistake if (i) less than 1% of the letters submitted by him are rejected and (ii) on 90% of days all the work submitted by him accepted.

11.35 250 passengers have made reservations for a flight from Delhi to Mumbai. The probability that a passenger who has reservation will not turn-up is 0.016. Find the probability that at the most 3 passengers will not turn-up.

11.36 The following mistakes per page were observed in a book. Fit a Poisson distribution to the data:

No. of mistakes per page	0	1	2	3	4
No. of pages	200	90	20	10	0

(Given: $e^{-.5} = .6065$)

11.37 A typist commits the following number of mistakes per page in typing 100 pages .Fit a Poisson distribution and calculate theoretical frequencies:

Mistakes per page	0	1	2	3	4	5
No. of pages	42	33	14	6	4	1

(given $e^{-.1} = .3679$)

11.38 Show that there is a inconsistency in the following statement: "The mean of a Poisson distribution is 16 and the standard deviation is 9".

11.39 Comment on the following statement: "For a Poisson distribution mean is 8 and variance is 7."

11.3 NORMAL DISTRIBUTION

The perfectly smooth and symmetrical curve, resulting from the expansion of the binomial $(q+p)^n$when n approaches infinitely is known as the normal curve. Thus, the normal curve may be considered as the limit toward which the binomial distribution approaches as n increases to infinity. Alternatively we can say that the normal curve represents a continues and infinite binomial distribution when neither p nor q is very small.

Even though p and q are not equal, for very large n we get perfectly smooth and symmetrical curve. Such curves are also called Normal Probability curve or Normal curve of Error or Normal Frequency curve.It was extensively developed and utilized by German Mathematician and Artsonomer Karl Gauss. This is why it called Gaussian curve in his honor.

The normal distribution was first discovered by the English Mathematician Dr. A. De Moivre. Later on French Mathematician Pierre S. Laplace developed this principle. This principle was also used and developed by Quetlet, Galton and Fisher.

11.3.1 Assumptions of Normal Distribution

The following are the four fundamental assumptions or conditions prevail among the factors affecting the individual events on which a normal distribution of observations is based:

1. *Independent Factors*: The forces affecting events must be independent of one another.
2. *Numerous Factors*: The causal forces must be numerous and of equal weight (importance).
3. *Symmetry* : The operation of the causal forces must be such that positive deviations from the mean are balanced as to magnitude and number by negative deviations from the mean.
4. *Homogeneity*: These forces must be the same over the population from which the observations are drawn (although their incidence will vary from event to event).

11.3.2 Equation of Normal Curve

We obtain the theoretical frequencies for X = 0,1,2,3,.........n using the expansion $N(q+p)^n$.if p = q then binomial distribution becomes a symmetrical frequency distribution. When n is very large it is very difficult to compute the expected frequencies. This difficulty is overcome by normal curve. Normal curve is a continuous algebraic curve. Its formula is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

or

$$y = \frac{Ni}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2}$$

where y = computed height of an ordinate of the curve at a given point of x-series i,e. at a distance of x from the mean.

N = number of items in frequency distribution,
i = width of the class-interval,
σ= Standard deviation of the distribution,

11.3.3 Characteristics or Properties of the Normal Curve/Normal Distribution

1. Bell shaped : The normal curve is a bell shaped curve and describes the probability distribution of a normal variate with two parameters μ (mean) and σ(standard deviation).
2. Continuous : Normal curve/Normal distribution is continuous.
3. Equality of central value : Being perfectly symmetrical values mean, median and mode of normal distribution are same. The ordinate of normal curve is maximum of the central value.
4. Unimodal : Since the maximum ordinate in normal curve is at one point only, hence normal distribution is unimodal, and normal curve is single humped.
5. Equal distance of Quartiles from median: The difference between third quartile and median $[Q_3 - M]$ is equal to the difference between median and first quartile $(M - Q_1)$.
6. Parameters and Constants : The two parameters of the normal distribution are mean (μ or X) and standard deviation (σ). The other constants are : Moments :

$$\mu_1 = 0, \mu_2 = \sigma^2, \mu_3 = 0, \mu_4 = 3\mu_2^2 = 3\sigma^4,$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0, \beta_2 = \frac{\mu_4}{\mu_3^2} = \frac{3\mu_2^2}{\mu_3^2} = 3,$$

$$\gamma_1 = 0, \gamma_2 = \beta_2 - 3 = 3 - 3 = 0$$

   Odd central moments are always zero.Skewness is zero and the curve is mesokurtic.

7. Points of Inflexion : Near the mean value, the normal curve is concave while near $\pm 3\sigma$it is convex to the horizontal axis. the points where the change in curvature occurs are $\pm 1\sigma$.

8. Area under the curve : The area lying between the normal curve and the horizontal axis is said to be the area under the curve and is equal to 1 in case of probabilities and is equal to total frequency in case of frequencies in the distribution.

   Based on mean and standard deviation the specific area under normal curve are as follows:

   (a) With in a range 0.6745 of σon the both sides the middle 50% of the observations occur i,e. mean $\pm 0.6745\sigma$ covers 50% area 25% on each side.

   (b) Mean $\pm 1 S.D.$ (i,e.$\mu \pm 1\sigma$) covers 68.268% area, 34.134 % area lies on either side of mean.

   (c) Mean $\pm 2 S.D.$ (i,e. $\mu \pm 2\sigma$) covers 95.45% area, 47.725% area lies on either side on mean.

   (d) Mean ± 3 $S.D.$ (i,e. $\mu \pm 3\sigma$) covers 99.73% area, 49.856% area lies on the either side of the mean.

   (e) Only 0.27% area is outside the range $\mu \pm 3\sigma$.

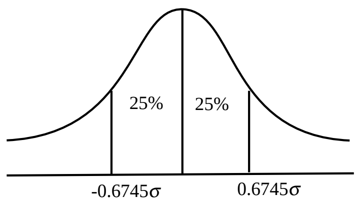


Figure 11.1. $\mu \pm 0.6745\sigma$

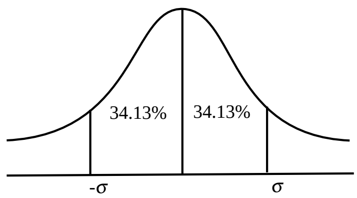


Figure 11.2. $\mu \pm \sigma$

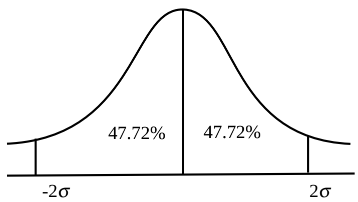


Figure 11.3. $\mu \pm 2\sigma$

11.3.4 Standard Normal Distribution

It is useful to transform a normally distributed variable into such a form that a single table of areas under the normal curve would be applicable regardless of the unity of the original distribution.

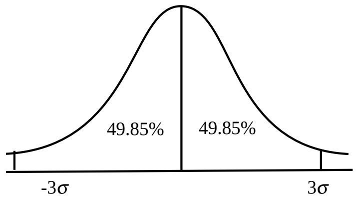


Figure 11.4. $\mu \pm 3\sigma$

Let X be random variable distributed normally with mean $\bar{X}$ and standard deviation σ, then we define a new random variable Z as

$$Z = \frac{X - \bar{X}}{\sigma}$$

where Z = New random variable, $\bar{X}$ =value of X, X = Mean of X, σ =S.D. of X. Then Z is called a standard normal variate and its distribution is called standard normal distribution with mean 0 and standard deviation unity. Similarly we can obtain the values of Z for any given value of X as shown in figure 3:

In the table "Area under Normal Curve" first column is Z in which the numbers are written from 0.0 to 3.0. In next, there are 10 columns representing the numbers 0.00 to 0.09 in which areas are given. Suppose the value of $Z = 1.45$. Then read the area corresponding 1.4 under 0.05 where it is 0.4265. Thus the area from 0 to 1.45 is 0.4265 or 42.65%. For positive Z the area is to the right of mean and for negative the area is to left to mean. To find the area from 1.45 to infinite subtract 0.4265 from 0.5000. Thus area to the right of 1.45 is 0.5000-0.4265 = 0.0735. The area between Z_1 = -1.45 and Z_2=1.05 is obtained as follows:

$$\begin{aligned} \text{area from } -1.45 \text{ to } 0 &= 0.4265 \\ \text{area from } 0 \text{ to } 1.05 &= 0.3531 \\ \therefore \text{required area} &= 0.4265 + 0.3531 = 0.7796 \end{aligned}$$

The area between Z_1= -1.45 and Z_2 = -1.05 is obtained as follows :

$$\begin{aligned} \text{area from } -1.45 \text{ to } 0 &= 0.4265 \\ \text{area from } -1.05 \text{ to } 0 &= 0.3531 \\ \therefore \text{required area} &= 0.4265 - 0.3531 = 0.0734 \end{aligned}$$

■ **EXAMPLE 11.15**

How would you use normal distribution to find approximately the frequency of exactly 5 successes in 100 trials, the probability of success in each trial being $p = 0.1$?

Solution:

Let the number of trials = n, then $n = 100,\ p = 0.1$ and $q = 1 - p = 1 - 0.1 = 0.9$.

Mean in case of binomial distribution

$$M = np = 100 \times 0.1 = 10$$

and

$$\sigma = \sqrt{npq} = \sqrt{100 \times 0.1 \times 0.9} = \sqrt{9} = 3$$

When n is large, n binomial distribution tends to normal distribution which is continues. Hence, the frequency of 5 success will be in the class $47.5-5.5$ and the mean and standard deviation of binomial distribution will be the mean and standard deviation of normal distribution.

$$\text{for } X = 4.5, Z_1 = \frac{4.5-10}{3} = -1.83$$

$$\text{for } X = 5.5, Z_2 = \frac{5.5-10}{3} = -1.5$$

$$\text{for } -1.83\sigma Area = 0.4664$$

$$-1.5\sigma Area = 0.4332$$

∴Area between -1.83 and -1.5 = 0.4664- 0.4332 = 0.0332. When total frequency is N, then expected frequency of the class $4.5-5.5$ is $0.332N$. Let $N = 100$, then expected frequency is 3.32%.

□

■ EXAMPLE 11.16

The mean of a distribution is 60 with a standard deviation of 10. Assuming that the distribution is normal, what percentage of items be (i) between 60 and 72, (ii) between 50 and 60, (iii) beyond 72 and (iv) between 70 and 80?

Solution:

(i) Between 60 and 72 :

$$Z = \frac{x-\mu}{\sigma} = \frac{72-60}{10} = 1.2$$

Area from 0 to 1.2 = 0.3849. ∴Percentage of observations between 60 to 72 = 0.3849 × 100 = 38.49%

[From the table the area between 0 and 1.26 = 0.3849]. Hence, the area between 60 and 72 = 0.3849 × 100 = 38.49%.]

(ii) Between 50 and 60:

$$Z_1 = \frac{x-\mu}{\sigma} = \frac{50-60}{10} = \frac{-10}{10} = -1$$

$$Z_2 = \frac{65-60}{10} = 0$$

Area between -1 and 0 = 0.3413

∴Percentage of observations = 0.3413 × 100 = 34.13%

(iii) Beyond 72:

Since half area is 0.5, hence area for more than 72 = 0.5 - 0.3849 = 0.1151

∴Percentage of observations more than 72 = 0.1151 × 100 = 11.51%

(iv) Between 70 and 80 :

$$Z_1 = \frac{70-60}{10} = \frac{10}{10} = 1$$

Area 0.3413

$$Z_2 = \frac{80 - 60}{10} = \frac{20}{10} = 2$$

Area 0.4772

Area between 70 and 80 = 0.4772 - 0.3413 = 0.1359

$\therefore$Percentage of observations between 70 and 80 = 0.1359 $\times$ 100 = 13.59%

□

■ EXAMPLE 11.17

In a normal distribution 31% of the items are under 45 and 8% are over 64. Find the mean and S.D. of the distribution.

Solution:

Let $\bar{X}$ = Mean of the distribution, σ= Standard Deviation of the distribution. Since the area to the left of the ordinate at $x = 45$ is 0.31, hence the area from the ordinate at $x = 45$ and Mean = $0.5 - 0.31 = 0.19$. The corresponding value of Z for this area from the table is - 0.5. Hence,

$$Z = \frac{45 - \bar{x}}{\sigma} = -0.5 \Rightarrow 45 - \bar{x} = -0.5\sigma$$

Since the area to the right of the ordinate at $x = 64$ is 0.08, hence the area between Mean and the ordinate at $x = 64 = 0.5 - 0.08 = 0.42$. Corresponding value of Z for this area from the table is 1.4. Hence

$$Z = \frac{64. - \bar{x}}{\sigma} = 1.4 \Rightarrow 64 - \bar{x} = 1.4\sigma$$

On solving both equations, we get

$$\sigma = 10, \ \bar{x} = 50$$

$\therefore$Mean, $\bar{x} = 50$ and standard deviation σ= 10.

□

■ EXAMPLE 11.18

The mean length of steel bars produced by a company is 10 meter and standard deviation 20 cm. 5000 bars are purchased by a building contractor.How many of these bars expected to be shorter than 9.75 meter in length? For your answer assume that the length of steel are normally distributed and use the following extract from table of areas from mean to distances ($\frac{X-\bar{X}}{\sigma}$) from mean under the normal curve :

$\frac{X-\bar{X}}{\sigma}$	1.10	1.15	1.20	1.25	1.30
Area	.3643	.3749	.3849	.3944	.4032

Solution:

Here, $\bar{X} = 10, \ \sigma = 0.2$

$$Z = \frac{(X - \bar{X})}{\sigma}, \ X = 9.75 \Rightarrow Z = \frac{9.75 - 10}{.2} = 1.25$$

Area from -1.25 to mean = 0.3944. Area to the left of the ordinate = 0.5000 - 0.3944 = 0.1056. The proportion of the bars which are shorter than 9.75 meter = 0.1056 and the number = 0.1056 $\times$ 5000 = 528

□

EXAMPLE 11.19

15000 students sat for an examination. The mean marks was 49 and the distribution of marks had a standard deviation of 6. Assuming that the marks were normally distributed what proportion of students scored (a) more than 55 marks, (b) more than 70 marks.

Solution:

(a) $Z = \frac{X-\bar{X}}{\sigma}$ and $X = 55, \bar{X} = 49, \sigma = 6$

$$Z = \frac{55-49}{6} = 1$$

The proportion of the students who got marks between 49 and 55 (i,e., 0 and 1 on standard scale)=0.3413

Hence the proportion of the students getting marks more than 55 = 0.5-0.3413 =0.1587

(b)

$$Z = \frac{X-\bar{X}}{\sigma} = \frac{70-49}{6} = \frac{21}{6} = 3.6$$

Area corresponding to $Z = 3.5$ is 0.4998.

$\therefore$ Proportion of students getting marks more than $70 = 0.5 - 0.4998 = 0.0002$ and the required number of students $= 0.0002 \times 15000 = 3$ □

EXAMPLE 11.20

The marks obtained in some degree examination are normally distributed , μ= 500 and $\sigma =$ 100 ; we have to pass 550 students out of 674 students who are appearing in the examination. What should be the minimum marks for passing the examination ?

Solution:

The proportion of successful students = $\frac{550}{674} = 0.8162$. From the normal table the value of Z corresponding to $0.8162 - 0.50 = 0.3162$ is 9. Hence the value of Z to the left of mean is -9. From $Z = \dfrac{X-\bar{X}}{\sigma}$, where X = Maximum marks for passing

$$-.9 = \frac{X-500}{100}$$

$$\Rightarrow X = -90 + 500 = 410$$

Thus minimum passing marks = 410

□

EXAMPLE 11.21

As a result tests on 20000 electric bulbs manufactured by a company it was found that the life time of the bulbs was normally distributed with an average of 2040 hours and a standard deviation of 60 hours . on the basis of this information estimate the number of bulbs that are expected to burn for (a) more than 2150 hours and (b) less than 1960 hours. the area, A (Area between 0 and Z) under normal curve is given:

Z	1.23	1.33	1.43	1.53	1.63	1.73	1.83
A	.3907	.4082	.4236	.4370	.4484	.4582	.4664

Solution:

Here

$$\bar{X} = 2040,\ \sigma = 1960,\ N = 20000$$

(a)

$$Z = \frac{X - \bar{X}}{\sigma} = \frac{2150 - 2040}{60} = \frac{110}{60} = 1.83$$

Area to the right of $1.83 = 0.5 - 0.4664 = 0.0336$

$\therefore$ Required Number = $0.0336 \times 20000 = 672$

(b)

$$Z = \frac{X - \bar{X}}{\sigma} = \frac{1960 - 2040}{60} = -1.33$$

Area to the left of $-1.33 = (0.5)(0.4082) = .0918$

$\therefore$ Required Number = $0.9018 \times 20000 = 1836$

□

Problems

11.40 1000 light bulbs with a mean life of 120 days are installed in a new factory. Their length of life is normally distributed with standard deviation 20 days. (i) How many will expire in less than 90 days ? (ii) It is decided to replace all the bulbs together what interval interval should be allowed between replacement if not more than 10% should expire before replacement .

11.41 In a distribution exactly normal 7 % of the items are under 35 and 89% are under 63. What are the mean and S.D. of the distribution ?

11.42 The mean height of the 1000 workers in a steel plant is 67 inch with a standard deviation of 5 inch.How many workers are expected to be above 72 inch in that plant?

11.43 If the height of 500 students are normally distributed with mean 65 inch and standard deviation 5 inch. How many students have height :

- **a)** greater than 70 inch.
- **b)** between 60 and 70 inch.

11.44 The Mumbai Municipal Corporation installed 2000 bulbs in the streets of Mumbai. If these bulbs have an average life of one thousand burning hours in with a S.D. of 200 hours, what numbers of bulbs might be expected to fail in the first 700 burning hours ? the table of the normal curve at selected values is as follows :

$\frac{X-\bar{X}}{\sigma}$	1	1.25	1.50
Probability	0.159	0.106	0.067

11.45 If the mean height of soldiers is 68.22 inch with a variance of 10.8 inch, how many soldiers in a regiment of 1000 can be expected to be over 6 feet all ? [In a normal distribution are to the right of an ordinate at 1.15 = 0.1251.]

11.46 The income per month of a group of 10000 persons was found to be normally distributed with mean Rs. 750 with a standard deviation of Rs. 50. show that of this group 95% had income exceeding Rs. 668 and only 5% had income exceeding Rs. 832. What was the lowest income among the richest 100?

11.47 In a statistical survey of 1000 small business firms in a city , it was found that their monthly average sales amounted to Rs. 8000 with a standard deviation of Rs. 2000. Assuming that the sales are normally distributed,estimate :

a) the number of firms whose monthly average sales were less than Rs. 6000 and

b) the number of firms whose monthly average sales were between Rs. 7000 and Rs. 9000. area under the normal curve

$Z = \frac{x-\mu}{\sigma}$	0.1	0.5	1.0	1.5	2.0
Area	0.0398	0.1915	0.3413	0.4332	0.4772

11.48 A sales-tax officer has reported that the average sales of the 500 business that he has to deal with during a year amount to Rs. 360000 with a standard deviation of Rs. 100000.Assuming that the sales in these business are normally distributed, find:

a) the number of business the sales of which are over Rs. 400000.

b) The percentage of business, the sales of which are likely to range between Rs. 300000 and Rs. 400000.

c) The probability that the sales of business selected at random will be over Rs. 300000. Area under the Normal Curve

$\frac{X-\mu}{\sigma}$	0.25	0.40	0.50	0.60
Area	0.0987	0.1554	0.1915	0.2257

11.49 In a normal distribution 7 percent of the items have values under 35 and 89 percent of the items have values under 62. Find the men and standard deviation of the distribution (In a standard normal distribution the area between means and x = 1.4757 is 0.43 and the area between mean and x = 1.2243 is 0.39).

CHAPTER 12

SAMPLING THEORY

12.1 UNIVERSE OF POPULATION

In any statistical investigation often called statistical survey, observations are made on a group of objects or individuals called elementary units as they are without any interference. The aggregate of individuals under study in any statistical survey is called a population.

According to A.C. Rosander

> A Population is the totality of objects under consideration.

In the words of Simpson and Kafka

> A universe or population may be defined as an aggregate of items possessing a common trait or traits.

According to G.Kalton,

> In statistical usage the term population does not necessarily refer to people but is a technical term used to describe the complete group of persons or objects for which the results are to apply.

12.1.1 Types of Population

There are two bases to classify populations:

1. Population Based on number of objects
    - Finite
    - Infinite

2. Population Based on existence of objects

   Real

   Hypothetical

12.1.1.1 Finite or Infinite Population A population is said to be finite if the number of individuals is fixed, i.e., finite. A population is said to be infinite if it is composed of infinitely large number of individuals. For examples.
Students of your university constitute a finite population, leaves of a tree constitute an infinite population (here infinite means very large).

12.1.1.2 Real or Hypothetical Population A population is said to be real, true or existent if it contains concrete and existing objects. For example, the workers in a factory, etc.

A population is said to be hypothetical or artificial if it is imaginary or it is constructed hypothetically on a paper by the statistician in his laboratory. This is done to illustrate certain statistical principles. For example, the possible outcomes of heads and tails in tossing a coin.

12.1.2 Sample

A part of the population selected to know some thing about the population is called a sample. The number of individuals selected in a sample is called its size.

According to G. W. Snedecore and W.G. Cochran,

> A sample consists of a small collection from a larger aggregates about which we seek information.

The statistical procedure of drawing a sample from the population is called sampling. the statistical procedure which are used for drawing inferences or conclusions about the population from the sample data are covered under inferential statistics or statistical inference. Thus sampling theory is the basis of statistical inference in which we wish to obtain maximum information about the population with minimum effort and maximum precession.

The main object of the study of sample or sampling is to get maximum information about the population under consideration at reduced cost, time and energy. According to weather burn,

> The theory of sampling is concerned first, with estimating the parameters of the population from those of the sample and secondly with gauzing the precision of the estimates.

12.2 CENSUS VS SAMPLING

In statistical investigation information about population is obtained in two ways:

1. Census method
2. Sampling method

Census Investigation:. A census investigation is one in which all the elementary units connected with the problem are studied. In other words, a process of investigation in which information is collected from each and every individuals of the population is called Census Method or Complete Enumeration. For example, population census conducted in India once in every ten years.

Sample Investigation:. A sample investigation is one in which only a selected group of individuals is studied (surveyed). In other words, the process of investigation in which information is obtained from a representative part of the population (called sample) is called sampling method or sample enumeration or sample survey.

12.3 PARAMETER AND STATISTIC

It is necessary to know the meaning of the two words 'Parameter' and 'Statistic' to study the sampling theory. According to Prof. R.A. Fisher, "Statistical measures of population, e.g., mean, standard deviation, correlation coefficient, etc. are called parameters or constants and the statistical measures obtained from sample values are called statistics." For example, The average or standard deviation or any other measure of the wages of 5,000 workers are parameters. But if we select 500 workers out of these 5,000 workers and compute average or standard deviation or any other measure of their wages, then they are statistics.

The fundamental problem of sampling theory is to study the relationship between the parameters of the population and statistics of the sample. In other words, "Sampling theory deals with the questions like. How the sample statistics (mean, standard deviation etc.) are reliable to the level of population parameters (mean, standard deviation etc.)? What is the difference between the statistical measures obtained from two samples? Is this difference due to chance or other reasons?

12.4 PRINCIPLES OF SAMPLING

The following are two important principles which determine the possibility of arriving at a valid statistical inference about the features of a population or process:

- Principle of statistical regularity
- Principle of inertia of large numbers

Principle of statistical Regularity. This principle is based on the mathematical theory of probability. According to Professor King

> The low of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to process the characteristic of the large group.

This Principle on emphasizes on two factors:

- Sample size should be Large
- Samples must be Drawn randomly:

Principle of Inertia of Large Numbers. This principle is a corollary of the principle of statistical regularity and plays a significant role in the sampling theory. This principle states that under similar conditions as the sample size get large enough the statistical inference is likely to be more accurate and stable. For example, if a coin is tossed a large number of times then relative frequency of occurrence of head and tail is expected to be equal.

12.5 STATISTICAL HYPOTHESIS

Any assumption regarding a population is called a *statistical hypothesis*. For example, 'Population mean is 50'. This statement about the population is a hypothesis.

To test this hypothesis whether it is true or not, on the basis of a sample is known *testing of hypothesis*. Statistical hypothesis which is tested under the assumption that it is true is called is null hypothesis. To accept a null hypothesis implies to reject some other complimentary hypothesis. This other hypothesis is known as *alternative hypothesis*.

12.5.1 Notation

If we wish to setup the hypothesis for the statement 'Population mean is 50', then we express the null hypothesis as

$$H_0 : \mu = 50$$

where, μ is population mean. Thus, alternative hypothesis may be:
Two-tailed alternative

$$H_1 : \mu \neq 50$$

Right-tailed alternative

$$H_1 : \mu > 50$$

Left-tailed alternative

$$H_1 : \mu < 50$$

12.5.2 Errors in hypothesis testing

There can be two types of error in hypothesis testing :

1. Error of type I: Null hypothesis is rejected when it is true.
2. Error of type II: Null hypothesis is accepted when it is not true.

These errors can be represented in tabular form as follow:

True Situation	Decision/Conclusion	
	H_0 accepted	H_0rejected
H_0 true	Correct decision	Type I error
H_0 not true	Type II error	Correct decision

The probability of committing type-I error is denoted by α, and the probability of committing type-II error is denoted by β.

12.5.3 Level of significance

The probability of making type-I error is termed as *level of significance*.

When we select a particular level for significance, say 5% and the probability of accepting a true hypothesis is 95%. As we reduce the level of significance, we reduce the probability of committing type I error. But note that the probability of committing type II error may increase. Thus, we can not reduce the probabilities of committing type I or type II error simultaneously. Usually, the level of significance is considered 5% or 1%.

12.5.4 Critical region

A region in the sample space S in which if the computed value of test statistic lies, we reject the null hypothesis, is called rejection region or critical region.

12.6 TESTS OF SIGNIFICANCE

Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favor or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as tests of significance

12.6.1 Procedure of Tests of Significance

Testing of hypothesis and test of significance are synonymous in some sense. On the basis of sample when a hypothesis of no difference is rejected we say that the difference between sample statistic and population parameter is significant. If this hypothesis is accepted, we say that the difference is not significant. The difference, if any, is due to sampling fluctuations. The following is the procedure (or steps) for test of significances.

1. Determination of the Problem: The first and important function for a test of significance is to determine the problem. In other words, it is necessary to know that in what context the decisions are to be made. These decisions may be to test the difference between statistic and parameter or to accept of reject a hypothesis.
2. Setting up of a Null Hypothesis: To know the significance of difference between statistic and parameter (or observed and expected) we have to set up a null hypothesis By null hypothesis, we have to set up a null hypothesis. By null hypothesis, we mean 'there is no difference between parameter and statistic;' and if any it is due to chance or due to fluctuations of sampling; For example,
   (a) the assumption, 'regular study to the students is not fruitful for success' and
   (b) the assumption, ' the new medicine will not cure'. are null hypothesis. A null hypothesis is denoted as H_0.
3. Selection of Level of Significance: A predetermined hypothesis is tested at some level of significance. In general, we use 10% (or 0.01) and 5% (or 0.05) level of significance.
4. Computation of Standard Error: The fourth step is to compute the standard error of the statistic to be used in the process. These are different formula for the standard error of various statistics which are used frequently. For example, Standard Error for mean, $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$, Standard Error for standard deviation $\sigma_\sigma = \frac{\sigma}{\sqrt{2n}}$, Standard Error for correlation coefficients, $\sigma_r = \frac{1-r^2}{\sqrt{n}}$
5. Computation of the Ratio Significance: To find a ratio significance (often called test statistic) are divide the difference between parameter and statistic by the corresponding standard error. For example,
$$\text{Ratio Significance for mean } = \frac{\overline{x}-\mu}{\sigma_x}$$
6. Interpretation: The last step of test of significance is interpretation (or to draw conclusion). For this we compare the predetermined critical value with the computed value. If computed value of ratio significance is more than critical value at 5% level of significance, i.e., if computed value > 1.96, we say that difference is significant. This means the difference is not due to fluctuations of sampling but there are other causes.

12.6.2 Large Sample and Small Sample

On the study of statistical research methods it is clear the sample size affects the reliability of sample information. The general assumption is that large sample is more reliable. So far test of significance purpose, we divide sample into two groups.

1. Large Samples
2. Small Samples

The basic assumptions in large sample and small sample are different that is why we use different methods of tests of significance. There is no definite line between large sample and small sample, but when sample size n is 30 or more ($n \leq 30$), it is a large sample and when $n < 30$ it is a small sample.

12.7 TESTS OF SIGNIFICANCE OF LARGE SAMPLES

12.7.1 Test for Binomial proportion

EXAMPLE 12.1

A coin is tossed 1,000 times and it falls towards the head 450 times. Is the coin not symmetrical? Is the coin biased?

Solution:

H_0 : The coin is symmetrical. Chance of getting head,

$$p = \frac{1}{2}, \quad q = 1 - p = \frac{1}{2}$$

$$\text{Standard Error} = \sqrt{npq} = \sqrt{1000 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{250} = 15.8$$

Expected number of heads in 1,000 tosses, $x = np = 1,000 \times 12 = 500$ test statistic

$$z = \frac{x - np}{\sigma_p} = \frac{500 - 450}{15.8} = 3.16$$

$\therefore$ The difference between observed number of heads and expected number of heads is greater than three times of the standard error; hence, the coin is not symmetrical.

□

EXAMPLE 12.2

400 heads and 100 tails resulted from 500 tosses of a coin. Find the limits for expected frequency of heads.

Solution:

Here

$$n = 500,\ x = 400,\ n - x = 100$$

Expected number of heads $= nP$, where P is unknown. So we have to estimate P by $\frac{x}{n}$ Thus

$$\hat{P} = p = \frac{x}{n} = \frac{400}{500} = \frac{4}{5}$$

where

$$x = nP = 400$$

Standard error of the number of heads x,

$$\begin{aligned}
\sigma_x &= \sqrt{nPQ} \\
&= \sqrt{npq} \text{ (Since P is not known, PQ is replaced by pq)} \\
&= \sqrt{500 \times \frac{4}{5} \times \frac{1}{5}} = \sqrt{80} = 8.94 \\
3\sigma_x &= 3 \times 8.94 = 26.82
\end{aligned}$$

$\therefore$Thus
Lower limit $= x - 3\sigma_x = 400 - 26.82 = 373.18 \approx 373$
Upper limit $= x + 3\sigma_x = 400 + 26.82 = 426.82 \approx 427$

□

EXAMPLE 12.3

An investigator reports 1,700 sons and 1,500 daughters. Do these figures confirm to the hypothesis that the sex ratio is $\frac{1}{2}$?

Solution:

$$\text{Total number of children} = 1{,}700 + 1{,}500 = 3{,}200$$

$$\text{No. of Sons} = 1{,}700$$

$$\therefore \text{Observed of sons proportion, } p = \tfrac{1700}{3200} = 0.53125$$

H_0 : Null hypothesis : The sons and daughters are in the same proportion.

$$P = \tfrac{1}{2} = 0.5 = \text{Expected proportion}$$

Standard Error

$$\sigma_p = \sqrt{npq} = \sqrt{\frac{1}{2} \times \frac{1}{2} \times 3200} = \sqrt{800}$$

test statistic

$$z = \frac{x - np}{\sigma_p} = \frac{1700 - 3200 \times \frac{1}{2}}{\sqrt{800}} = 3.5355$$

Since the difference is greater than 3 times of the standard error, the sex ratio is not $\frac{1}{2}$. □

EXAMPLE 12.4

It is desired to estimate the proportion of people in a certain town who have been vaccinated against small pox. Out of a random sample of 600 only 150 had been vaccinated. Find also the 95% confidence limits for the population.

Solution: Number of people in the sample, $n = 600$
Proportion of people who have been vaccinated, $\frac{a}{n} = p = \frac{150}{600} = \frac{1}{4}$ or 0.25
Proportion of people who have not been vaccinated, $q = 1 - 0.25 = 0.75$
Standard Error of the proportion,

$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.25 \times 0.75}{600}} = 0.0177$$

95% confidence limits for the proportion:

$$p \pm 1.96\sqrt{\frac{pq}{n}} \Rightarrow 0.25 \pm 1.96 \times 0.0177$$

and
Upper limit = 0.25 + 0.0347 = 0.2847
Lower limit = 0.25 - 0.0347 = 0.2153
i.e., 0.215 and 0.285 or 21.5 percent and 28.5 percent. □

EXAMPLE 12.5

A random sample of 500 apples were taken from a large consignment and 65 were found to be bad. Estimate the proportion of the bad apples in the consignment, as well as the standard error of the estimate. Deduce that the percentage of bad apples in the consignment almost certainly lies between 8.5 and 17.5

Solution: Proportion bad apples,

$$p = \frac{65}{100} = 0.13$$

and

$$q = 1 - 0.13 = 0.87$$

Standard Error of proportion,

$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.13 \times 0.87}{500}} = 0.015$$

Confidence limits for proportion :

$$= p \pm 3\sigma_p \times 100$$

$$= 0.13 \pm 3 \times 0.015$$

and

$$= 0.13 \pm 0.045$$

Lower limit = 13-0.045 = 0.085 and upper limit = 0.13 + 0.045 = 0.175. In percentage, $0.85 \times 100 =$ 8.5% to $0.175 \times 100 = 17.5\%$. □

12.7.2 Standard Error and Sample Size

We know that standard error of proportion $\sigma_p = \sqrt{\frac{pq}{n}}$ depends the value of p and the sample size n. Thus the limits of p does not dependent on population size. If p remains constant, and n change, then standard error of p will also change. Mathematically, the value of standard error is inversely proportion to the under root of n.

12.7.3 Test for the difference between Proportion of Two Samples

Suppose two random samples of size n_1 and n_2 are drawn from a population. Let p_1 and p_2 be the sample proportions respectively for a certain attribute. To test the null hypothesis. The process of the test is as follows :

1. Null Hypothesis : There is no difference in the two proportions or the two samples have been taken from the same population.
2. To estimate the parameter, proportion : We estimate the combined proportion by

$$p_0 = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}, q_0 = 1 - p_0$$

   where p_0 is parameter proportion.
3. Standard Error of the difference of proportions
   Standard Error for $p_1 = \sqrt{\frac{p_0 q_0}{n_1}}$
   Standard Error for $p_2 = \sqrt{\frac{p_0 q_0}{n_2}}$
   Standard Error for difference $p_1 \sim p_2$

$$S.E._{p_1 \sim p_2} = \sqrt{p_0 q_0 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

   Interpretation: If the difference between p_1 and p_2 is greater than 3 times the standard error of the difference then the difference is significant otherwise not significant.

■ EXAMPLE 12.6

In a random sample of 500 mean from a particular district of U.P. 300 are found to be smokers. Out of 1,000 men from another district 550 are smokers. Do the data indicate that the two districts are significantly different with respect to the prevalence of smoking among men?

Solution:

Proportion of smokers in the two districts together

$$p_0 = \frac{300+500}{500+1,000} = \frac{850}{1,500} = 0.566666 \approx 0.567$$

hence

$$q_0 = 1 - 0.567 = 0.433$$

Here $p_1 = \dfrac{300}{500} = 0.6$ and $p_2 = \dfrac{550}{1000} = 0.55$

$$\begin{aligned} S.E._{p_1 \sim p_2} &= \sqrt{p_0 q_0 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\ &= \sqrt{0.567 \times 0.433 \left(\frac{1}{500} + \frac{1}{1000}\right)} \\ &= 0.027 \end{aligned}$$

Since the difference $p_1 \sim p_2$ is less than twice of the standard error, the two districts are not significantly different with respect to the prevalence of smoking among men.

□

■ EXAMPLE 12.7

In a consumers preference survey a simple sampling method 60% of the persons gave preference in city A for a certain commodity. In city B the similar percentage was 50%. If the sample selected in the first case was of 600 and in the second case it was 500, what are your conclusions regarding teh difference in preference of the consumers in two cities?

Solution: H_0 : Null Hypothesis : There is no difference in preference of the consumers in two cities.

Proportion of City A (p_1) = 60% or 0.60

Proportion of City B (p_2) = 60% or 0.50

Combined Proportion,

$$\begin{aligned} p_0 &= \frac{(0.6 \times 600) + (0.5 \times 500)}{600 + 500} \\ &\approx 0.5545 \\ q_0 &\approx 0.4455 \end{aligned}$$

$$\begin{aligned} S.E._{p_1 \sim p_2} &= \sqrt{p_0 q_0 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\ &= \sqrt{0.5545 \times 0.4455 \left(\frac{1}{600} + \frac{1}{500}\right)} \\ &= 0.03009 \end{aligned}$$

$$z = \frac{p_1 \sim p_2}{S.E._{p_1 \sim p_2}} = \frac{0.60 - 0.50}{0.03009} = 3.32$$

Since the difference $p_1 \sim p_2$ is more than three times of the standard error, the difference in preference is significant. □

12.7.4 Standard Error when Sample Proportion is not equal

Let P_1 and P_2 be the two population proportion for an attribute A. Let p_1 and p_2 be the two sample proportions from these two populations respectively. We are interested in the problem, is by increasing the sample size the difference between P_1 and P_1 be hidden. In this case:

Standard Error for p_1,

$$S.E._{p_1} = \sqrt{\frac{p_1 q_1}{n_1}}$$

Standard Error for p_2,

$$S.E._{p_1} = \sqrt{\frac{p_2 q_2}{n_2}}$$

Standard Error for $p_1 - p_2$

$$S.E._{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

As usual if the difference between p_1 and p_2 is greater than 3 times of the $S.E._{p_1 - p_2}$, the difference is significant otherwise not significant.

■ **EXAMPLE 12.8**

In a random sample of 1,000 blind persons 600 are found to be literate. In another similar sized sample of non-blind persons the number of literate persons was found to be 800. Is this difference significant.

Solution: Here,

$$p_1 = \frac{600}{1,000} = 0.6; q_1 = 1 - 0.6 = 0.4$$
$$p_2 = \frac{800}{1,000} = 0.8; q_2 = 1 - 0.8 = 0.2$$

$$S.E._{p_1 \sim p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$
$$= \sqrt{\frac{0.6 \times 0.4}{1000} + \frac{0.8 \times 0.2}{1000}}$$
$$= 0.02$$

$$z = \frac{p_1 \sim p_2}{S.E._{p_1 \sim p_2}} = \frac{0.6 \sim 0.8}{0.02} = 10$$

Since the difference is enough large by three times of the standard error, the observed difference is significant. □

■ **EXAMPLE 12.9**

In two large populations, there are 30% and 25% respectively of gray-colored people. Is this difference likely to be hidden at 5% level of significance in samples of 1,200 and 900 respectively from two population?

Solution:

$$p_1 = \tfrac{30}{100} = .3;\ p_2 = \tfrac{25}{100} = .25;$$

$$q_1 = 1 - .3 = .7;\ q_2 = 1 - .25 = .75$$

$$n_1 = 1,200\ n_2 = 900$$

Since the two samples are taken from different populations, the standard error of the difference is given by

$$\sigma_{p_1-p_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

$$= \sqrt{\frac{0.3 \times 0.7}{1,200} + \frac{0.25 \times 0.75}{900}} = 0.196$$

$$\frac{p_1 - p_2}{\sigma_{p_1-p_2}} = \frac{0.015}{0.196} = 2.55 < 3$$

Since difference is less than 3 times of standard error, hence the difference is likely to be hidden. If level of significance is taken as 5%, then $2.55 > 1.96$, hence difference is significant and it is not likely that the difference will be hidden. □

12.7.5 Test for Mean

12.7.5.1 Standard Error of Mean

1. When population standard deviation σ is known, then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

2. When population standard deviation is not known, then $\sigma_{\bar{x}} = \frac{S}{\sqrt{n-1}}$ or $\frac{S}{\sqrt{n}}$, where S is standard deviation of sample values.

■ EXAMPLE 12.10

A sample of 100 units is found to have 5 lb. as mean. Could it be regarded as simple sample from a large population where mean is 5.64 lb. and σ = 1.5 lb.

Solution:

Standard Error of Mean = $\frac{\sigma}{\sqrt{n}}$ = $\frac{1.5}{\sqrt{100}}$ = 0.15

Difference between population and sample means = $5.64 - 5 = 0.64$

Three times Standard Error of mean = $3 \times 0.15 = 0.45$

Since the difference is greater than three times of standard error, the sample could not be regarded as taken from this large population.

□

■ EXAMPLE 12.11

Given the following information about a random sample of individual items, estimate with 0.27 percent probability level the limits within which the population mean lies: Mean = 172, S.D. = 12, n=64

Solution:

Standard Error of Mean = $\frac{\text{Sample Standard Deviation}}{\sqrt{n}}$ = $\frac{12}{\sqrt{64}}$ = 1.5

Limits for population mean at 0.27% level of significance are : $172 \pm (3 \times 1.5)$, i.e., between 167.5 and 176.5

□

■ EXAMPLE 12.12

Find out n to be included in a sample from a universe with mean 100 and standard deviation 10 to ensure that the mean of the sample in all probability would be within 0.01 percent of the true value.

Solution:

We know that 99.99% values lie between mean ± 3.9 standard error i.e., 0.01% values are outside this range. Hence, then confidence interval is $\pm 3.9\sigma$.

$$|\text{Sample mean} - \text{Actual Mean}| \leq 0.01\% \times \text{population mean}$$
$$\leq \frac{0.01}{100} \times 100 = 0.01$$

Sampling error = Standard error $\times$ Critical Value

$$= \frac{10}{\sqrt{n}} \times 3.9 = \frac{39}{\sqrt{n}}$$

Taking this sampling error as 0.01, we have

$$0.01 = \tfrac{39}{\sqrt{n}} \text{ and } 0.01\sqrt{n} = 39$$

$$\Rightarrow n = (3900)^2 = 15210000$$

□

■ EXAMPLE 12.13

From a normally distribution infinite number of iron bars with mean and standard deviation as 4 ft. and 0.6 ft. respectively, a sample of 100 bars is taken. If the sample mean is 4.2 ft. can the sample be called a truly random sample?

Solution:

$$n = 100 \ \bar{X} = 4.2 \ \mu = 4 \ \sigma_p = 0.6$$

$$\text{Standard error of mean, } \sigma_{\bar{x}} = \tfrac{\sigma_p}{\sqrt{n}} = \tfrac{0.6}{\sqrt{100}} = 0.06$$

Now

$$\frac{\text{Difference}}{S.E.} = \frac{|X - \mu|}{\sigma_{\bar{x}}} = \frac{|4.2 - 4|}{0.06} = 3.33 < 3$$

Since difference is less than three times of the S.E., hence the sample can not be called a truly random sample.

□

■ EXAMPLE 12.14

If it costs a rupee to draw one number of sample, how much would it cost in sampling from a universe with mean 100 and standard deviation 100 and standard deviation 10 to take sufficient number as to ensure that the mean of the sample would in a 5% probability be within 0.01 percent of the true value. Find the extra cost necessary to double this precision.

Solution:

$$\bar{X} = 100; \sigma = 10$$

Difference between population mean and sample mean

$$= 0.01\% \text{ of population mean} = \tfrac{0.01}{100} \times 100 = 0.01$$

Standard error of mean = $\frac{\sigma_p}{\sqrt{n}} = \frac{10}{\sqrt{n}}$
The difference at 5% level of significance =

$$1.96 \times S.E. = \frac{10}{\sqrt{n}}$$

That is

$$\frac{1.96 \times 10}{\sqrt{n}} = 0.01$$

$$\Rightarrow n = (1,960)^2 = 38,41,600$$

The cost of selecting 38,41,600 units = Rs. 38,41,600 . By doubling the precision, the difference will be half or the standard error will be double.

$$(\frac{0.01}{2}) = 0.005 \left[\text{or} S.E. = \frac{10}{\sqrt{n}} \times 2 = \frac{20}{\sqrt{n}}\right]$$

$$\frac{1.96 \times 10}{\sqrt{n}} = 0.005$$

$$\Rightarrow n = (3,920)^2 = 1,53,66,400$$

The cost of selecting 1,53,66,400 units = Rs. 1,15,24,800

□

12.7.6 Test for the difference of two Sample Means

Suppose we draw two samples from two population (or some population). Then to test the hypothesis, 'The difference of the means of two samples is not significant or the two samples have been taken from the same population. For this we consider the difference between two sample means and the standard error of this difference.

As usual, if the difference is greater than three times standard error, its considered significant otherwise not significant.

1. When population standard deviation (σ_{pop}) is known:

$$S.E._{\overline{x_1}-\overline{x_2}} == \sqrt{\sigma_{pop}^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sigma_{pop} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

   where n_1 and n_2 are sample sizes.

2. When population standard deviation is not known:

$$S.E._{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_1}}$$

   where $\sigma_1$is S.D. of first sample and $\sigma_2$is S.D. of second sample.

3. When a sample mean $\bar{x_1}$ is compared with the combined mean of two sample, $(\overline{x_{12}})$:

$$S.E._{\bar{x_1}-\bar{x_2}} = \sqrt{\sigma_{pop}^2 \times \frac{n_2}{n_1(n_1+n_2)}} = \sigma_{pop} \cdot \sqrt{\frac{n_2}{n_1(n_1+n_2)}}$$

4. When the two series are correlated. That is the samples are taken from correlated populations.

$$S.E._{\bar{x_1}-\bar{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2r\frac{\sigma_1 \times \sigma_2}{n_1 \times n_2}}$$

$$\text{Test Statistic} = \frac{\text{Difference}}{S.E.}$$

■ EXAMPLE 12.15

The mean produce of wheat of a sample of 100 fields comes to 200 lb. per acre with a standard deviation of 10 lb. Another sample of 150 fields gives the mean at 220 lb. with a standard deviation of 12 lb. Assuming the standard deviation of yield is 11 lb. for the universe, find out if there is a significant difference between the yield of the two samples.

Solution:

Null hypothesis, H_0: There is no difference between the yield of two samples.

$$S.E._{\bar{x_1}-\bar{x_2}} = \sqrt{\sigma_{pop}^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\sqrt{(11)^2\left(\frac{1}{100} + \frac{1}{150}\right)} = \sqrt{121\left(\frac{1}{60}\right)} = 1.42$$

Observed difference = 220 - 200 = 20 lb. Since the difference is about 14 times $\left[\because \frac{20}{1.42} = 14\right]$ of the standard error, the two mean yields are significantly different.

□

■ EXAMPLE 12.16

A sample of 500 persons of a city showed a mean income of Rs. 120 per month with standard deviation of Rs. 10. Another sample of 1,000 persons a mean income of Rs. 123 per month with σ of Rs. 12, is the difference between the two means statistically significant?

Solution:

$$S.E._{\bar{x_1}-\bar{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\sqrt{\frac{10^2}{500} + \frac{12^2}{1000}} = \sqrt{\frac{100}{500} + \frac{144}{1000}} = \sqrt{\frac{344}{1000}} = 0.59$$

Now

$$\frac{\text{Difference}}{\text{Standard Error}} = \frac{|123 - 120|}{0.59} = 5.08$$

The difference is greater than 3 times of the standard error; hence it is significant.

□

■ EXAMPLE 12.17

A random sample of 100 villages in Agra district gives the mean population of 480 persons per village. Another sample of 150 villages from the same district gives the mean at 490. If the standard deviation of the mean population of villages in the district is 10, find out if the mean of the first sample is significantly different from the combined mean of the two samples taken together.

Solution:

$$\text{Combined Mean} = \frac{n_1\bar{x_1}+n_2\bar{x_2}}{n_1+n_2}$$

=

$$\frac{(100\times 480)+(150\times 490)}{100+150} = 486$$

Standard Error of difference between the mean of first sample and the combined mean of the samples.

$$= \sigma_{pop}\sqrt{\frac{n_2}{n_1(n_1+n_2)}}$$

=

$$10\sqrt{\frac{150}{100(100+150)}} = 0.774$$

Since the difference is 7.8 times $\left(\because \frac{6}{0.774} = 7.8\right)$ of the standard error; hence it is significant. □

■ EXAMPLE 12.18

The means of simple samples of 1,000 and 2,000 are 69.5 and 70 respectively. Can the samples be regarded as drawn from the same universe having a standard deviation 2?

Solution:

Difference of the two sample means $= 70 - 69.5 = 0.5$

Standard Error

$$S.E._{\bar{x_1}-\bar{x_2}} = \sigma_{pop}\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$3\sqrt{\frac{1}{1000}+\frac{1}{2000}} = 0.117$$

Now

$$\frac{\text{Difference}}{\text{Standard Error}} = \frac{0.5}{0.117} = 4.27 > 3$$

The difference is greater than 3 times of the standard error, hence it is significant i.e., the two samples have not been taken from same population. □

12.7.7 Standard Error of the difference between two Sample Medians

$$S.E._{M_1-M_2} = \sqrt{(S.E._{M_1})^2+(S.E._{M_2})^2}$$

■ EXAMPLE 12.19

Two samples of 100 and 80 students are taken with a view to find out their average monthly expenditure. It is found that median monthly expenditure for the first group is Rs. 85 and for the second group is Rs. 100. The standard deviation for the first group is Rs. 7 and for the second Rs. 8. Examine if the difference between the medians of the two samples is statistically significant.

Solution:

Given $n_1 = 100$, $M_1 = 85$, $\sigma_1 = 7$, $n_2 = 80$, $M_2 = 100$, $\sigma_2 = 8$

$$S.E._{M_1} = 1.25331\frac{\sigma}{\sqrt{n}}$$

$$S.E._{M_1} \text{for I sample} = 1.25331 \frac{7}{\sqrt{100}} = 0.877$$
$$S.E._{M_2} \text{for II sample} = 1.25331 \frac{8}{\sqrt{80}} = 2.121$$
$$S.E._{M_1 - M_2} = \sqrt{S.E._{M_1}^2 + S.E._{M_2}^2}$$
$$S.E._{M_1 - M_2} = \sqrt{0.877^2 + 1.121^2} = 1.42$$

Difference of the two medians = $100 - 85 = 15$, which is $15 + 1.42 = 10.56$ times of the standard error; hence the difference is statistically significant.

□

12.7.8 Standard Errors of difference between two sample Standard Deviations

1. When population S.D. is known

$$S.E._{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_{pop}^2}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{\sigma_{pop}}{\sqrt{2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

2. When population standard deviation is not known: Standard error of

$$\sigma_1 - \sigma_2 = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}} = \frac{1}{\sqrt{2}} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

3. When standard deviation of one sample is compared with the combined standard deviation of the two samples:

$$\text{Standard error of } \sigma_1 - \sigma_{1.2} = \sqrt{\frac{\sigma_{pop}^2}{2} \times \frac{n_2}{n_1(n_1 + n_2)}} = \frac{\sigma_{pop}}{\sqrt{2}} \times \sqrt{\frac{n_2}{n_1(n_1 + n_2)}}$$

■ **EXAMPLE 12.20**

The standard deviation in two samples of the sizes of 100 and 120 respectively were found to be 13.9 and 17.6 respectively. If population standard deviation is 15, find out if there is any significant difference between two standard deviation.

Solution:

$$\text{S.E. of } \sigma_1 - \sigma_2 = \sqrt{\frac{\sigma_{pop}^2}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$
$$= \sqrt{\frac{15^2}{2}\left(\frac{1}{100} + \frac{1}{120}\right)} = 1.436$$

Observed difference $(17.6 - 13.9) = 3.7$. Since the difference is 2.58 $[(3.7 \div 1.436) = 2.58]$ times of the standard error, hence it is not significant.

□

■ **EXAMPLE 12.21**

From the following data, test the significance of the difference of the standard deviations :

$$n_1 = 500; n_2 = 1000; \sigma_1 = 10; \sigma_2 = 12$$

Solution:

$$S.E_{\cdot\sigma_1-\sigma_2} = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}} = \sqrt{\frac{10^2}{2 \times 500} + \frac{12^2}{2 \times 1,000}} = 0.415$$

Difference (12-10) = 2, Since the difference is 5 times $\left[\frac{2}{0.415} = 5\right]$ of the standard error; hence it is significant.

□

■ EXAMPLE 12.22

The standard deviation of the wages of 800 workers of carpet industry is Rs. 25.6, another sample of 1,200 workers gives the standard deviation at Rs. 31.0. Find out if the standard deviation of the first sample differs significantly from the combined standard deviation of the two samples.

Solution:

Assuming that the mean of two samples are equal.
Combined standard deviation =

$$= \sqrt{\frac{(n_1 \times \sigma_1^2) + (n_2 \times \sigma 2^2)}{n_1 + n_2}}$$

$$= \sqrt{\frac{(800 \times 25.6^2) + (1,200 \times 31^2)}{800 + 1200}} = 28.96$$

$$S.E_{\cdot\sigma_1-\sigma_{1.2}} = \sqrt{\frac{\sigma_{pop}^2}{2} \times \frac{n_2}{n_1(n_1 + n_2)}}$$

$$= \sqrt{\frac{838.744}{2} \times \frac{1,200}{800(800 + 1,200)}} = 0.56$$

The difference between standard deviations of first sample and combined sample $28.96 - 225.6 = Rs.3.36$, is about 6 times $\left[\frac{3.36}{0.56} = 6\right]$ of the standard error hence significant.

□

Problems

12.1 160 heads and 240 tails resulted from 400 tosses of a coin. Find a 95% confidence interval for the probability of a head. Does this appear to be a true coin?

12.2 In a newspaper article of 1,600 words in Hindi, 64% of the words are found to be of Sanskrit origin. Assuming the simple sampling conditions hold good, estimate the proportion of Sanskrit words in the writer's vocabulary and assign limits to that proportion.

12.3 In 324 throws of six-faced die odd points appeared 181 times. Would you say that the die is fair? State carefully the property on which you base your calculation.

12.4 In a random sample of 1,000 persons from town A, 400 are found to be consumers of wheat. In a sample of 800 from town B, 400 are found to be consumers of wheat. Do these data reveal a significant difference between town A and town B, so far as the proportion of wheat consumers is concerned.

12.5 500 articles were selected at random out of a batch containing 10,000 articles and 30 were found to be defective. How many defective articles would you reasonably except to find in the whole batch?

12.6 In a town, 19,400 persons were observed and 27% of them were found to be deaf and in town B, 30% of the 29,750 persons observed were found to be deaf. Can the difference observed in the percentage of deaf persons be attributed solely to fluctuations of sampling?

12.7 In a year there were 956 births in town A of which 52.5% were male, while in town A and B combined, the proportion in a total of 1,406 births, males was 49.6%. Is there any significant difference between the proportion of births in two towns.

12.8 A random sample of 200 measurements from an infinite population gave mean value of 50 and standard deviation of 9. Determine the 95% confidence interval for the mean value of the population.

12.9 The means of simple samples of 1,000 and 2,000 are 67.5 are 68.0 inch respectively. Can the samples be regarded as drawn from the same population having a S.d. of 2.5?

12.10 A random sample of 200 villages was taken from Kanpur district and the average population per village was found to be 420 with an S.D. of 50. Another random sample of 200 villages from the same district gave a average population of 480 per village with an S.D. of 60. Is the difference between the averages of the two samples statistically significant?

12.11 Out of 10,000 students in a University, a sample of 400 is taken; the monthly average of the expenditure of these 400 students is found to be Rs. 75 and the standard deviation Rs. 2.50. Set the limits within which the average of additional samples would lie.

12.12 The mean produce of wheat of a sample of 100 fields is 200 kilogram per acre with a standard deviation of 10 kilogram. Another sample of 150 fields gives the mean at 220 kilogram with a standard deviation of 12 kilogram. Assuming the standard deviation of the mean field at 11 kilogram for the universe find at 1% level if the two resuls are consistent.

12.8 TEST OF SIGNIFICANCE FOR SMALL SAMPLES

The basis for the analysis of large samples is not applicable in case of small samples; specially when the number items is less than 30 (some persons take this number as 60). The assumptions for the analysis of large samples are not true in case of small samples. In large samples we can use the sample variance to find the standard error of mean when population standard deviation is not known. But in case of small sample the sample variance is not a reliable estimate of population standard deviation. Small sample do not conform to the law of inertia of large numbers. Also the interval for population mean given by $\pm 1.96\sigma\bar{x}$ is not true in case of small samples.

The problem is why do we take small samples? In the study of economic and social phenomena, we generally take large samples so the techniques of large sample test of significance can be used. But when data is obtained in laboratory or by experiments which requires a large expenditure and more time; then samples of large size are not considered good. Sometimes it is not possible to draw large samples.

Thus it becomes necessary to have a method which can deal with small samples

Bessele's Corrections:. In case of small samples standard error of mean is estimated by

$$\sigma_p = \sqrt{\left(\frac{n}{n-1}\right)\sigma_s^2}$$

Where $\left(\frac{n}{n-1}\right)$ is Bessels's correction

$$\therefore \sigma_p = \sqrt{\left(\frac{n}{n-1}\right)\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma d^2}{n-1}}$$

12.8.1 Test of Significance based on t-Distribution

One of the most important test of significance in case of small samples is t-test which is based on t-distribution was discovered by William Gosset under the pen name 'Student'. To compute a t-statistic we decide the difference of sample mean and population by the estimated standard error of mean.

We apply t-test to test the hypothesis μ =a certain value when population is normal, population standard derivation is not known and sample size is less than 30.

Here,

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n}} \text{ or } t = \frac{x-\mu}{S'/\sqrt{(n-1)}}$$

Here, $\bar{x}$ = sample mean, μ = population mean, S = sample standard deviation = $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ or $S' = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$, and n= sample size.

12.8.2 Test for the Mean of Small Sample

To test the significance between the difference of sample mean and population mean, we proceed as follows:

1. Null Hypothesis: The population mean is definite, say mean = 50 cm
2. Estimation of σ of Universe : For this we use the formula

$$S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \text{ or } S = \sqrt{\frac{\Sigma d^2}{n-1}}$$

3. Calculation of t-Statistic

$$t = \frac{\bar{x}-\mu}{S}\sqrt{n}$$

or

$$t = \frac{\bar{x}-\mu}{S'}\sqrt{n-1}$$

where

$$S' = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

We do not use t-test when population variance is known.

4. Determination of Degree of Freedom:

$$\text{Degree of Freedom } = \nu = n-K$$

where n = sample size, K = No. of constraints. For t-test degrees of freedom is (n-1).

5. Table value of t: For relevant degrees of freedom (d.f.) at given level of significance we note the table value of t. For example, for 8 d.f. at 5% level of significance the table value (from t-table) is 2.306.
6. Interpretation: Compare the calculate value of t with table value of t. If calculated value of t is greater than table value of t, then difference is significant and over null hypothesis is false.

■ EXAMPLE 12.23

Six boys are selected at random from a school and their marks in Mathematics are found to be 63,64,66,60,68 out of 100. In the light of these marks, discuss the general observations that the mean marks in Mathematics in the school were 66.

Solution: Null hypothesis, H_0 : Population mean, μ = 66

Marks (x)	Deviation from Mean $d = (x - \bar{x})$	d^2
63	-1	1
63	-1	1
64	0	0
66	2	4
60	-4	16
68	4	16
384	0	38

Sample mean, $\bar{x} = \frac{\Sigma x}{n} = \frac{384}{6} = 64$
Standard deviation of the sample

$$S = \sqrt{\frac{\Sigma d^2}{n-1}} = \sqrt{\frac{38}{6-1}} = 2.756$$

$$t = \frac{|\bar{x} - \mu|}{S}\sqrt{n} = \frac{66-64}{2.756} \times \sqrt{6} = 1.777$$

$$d.f. = (n-1) = (6-1) = 5$$

Computed value of $t = 1.777$ is less than value with 5 degrees of freedom at 5% level of significance; hence the null hypothesis 'mean = 66' is true. Thus we can say the average of marks in the school is 66

□

■ EXAMPLE 12.24

A sample of size 10 has Mean as 57 and standard deviation as 16. Can it come from a population with Mean 50?

Solution: Null Hypothesis : The sample is taken from population whose mean is 50

$$t = \frac{\bar{x} - \mu}{S}\sqrt{n-1} = \frac{57-50}{16}\sqrt{10-1} = 1.3125$$

$$d.f. = 10 - 1 = 9$$

For 9 d.f. at 5% level of significance the value of t=2.262
Computed value of $t = 1.3125 < 2.262$
Table value of t corresponding to 9 degrees of freedom at 5% level of significance is 2.262. The calculated value of t=1.3125 is considerable less than the table value. Hence, the hypothesis is not disproved and the sample may come from this population having mean 50.

Alternative Method . We are given sample standard deviation = 16. Previously we assumed it S' = $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$ = 16 and used the formula:

$$t = \frac{\bar{x} - \mu}{S'}\sqrt{n-1}$$

But if we assume it S=$\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = 16$, then use the formula:

$$t = \frac{\bar{x}-\mu}{S}\sqrt{n}$$

Thus

$$t = \frac{57-50}{16}\sqrt{10} = 1.38$$

□

Fixing Limits of Population Mean . The confidence limits for population mean are as follows:

95% confidence interval for $\mu = \bar{x} \pm \frac{S}{\sqrt{n}} t_{.05}$

99% confidence interval for $\mu = \bar{x} \pm \frac{S}{\sqrt{n}} t_{.01}$

■ **EXAMPLE 12.25**

A random sample of size 20 has 52 as mean and the sum of the squares of the deviations taken from mean is 171. Can this sample be regarded as taken from the population having 55 as mean? Obtain (a) 95%, and (b) 99% confidence limits of the mean of the population.

Solution: Given : $\bar{x} = 52, \mu = 55, n = 20$

Null hypothesis : $\mu = 55$

$$S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

$$\therefore S = \sqrt{\frac{171}{20-1}} = 3$$

$$t = \frac{\bar{x}-\mu}{S}\sqrt{n} = \frac{52-55}{3}\sqrt{20} = 4.47$$

For 19 d.f. at 5% level of significance, $t_{.05} = 2.093$

Computed value of t=4.47 ¿ 2.093

Hence, the null hypothesis is rejected. That is the sample is not taken from the population whose mean is 55.

Confidence Limits of the Population Mean . On 95% Confidence Interval: For 19 d.f., $t_{.05}$ = 2.093

$$\bar{x} \pm \frac{S}{\sqrt{n}} \times t_{.05} = 52 \pm \frac{3}{\sqrt{20}} \times 2.093 = 52 \pm 1.40$$

50.60 and 53.40

□

12.8.3 Test for the Difference Between Two Sample Means

To test the significance between the difference of two sample means taken from two populations with same variance (unknown), t is defined as follows:

$$t = \frac{\bar{x_1} - \bar{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{x_1} - \bar{x_2}}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Here $\bar{x_1}$ and $\bar{x_2}$ are respectively the mean of the two samples; n_1 and n_2 are the sample sizes and S^2 the combined variance of the two samples is given by:

$$S = \sqrt{\frac{\Sigma d_1^2 + \Sigma d_2^2}{n_1 + n_2 - 2}}$$

Where d_1 and d_2 are the deviations about the respective means. If S_1^2 and S_2^2 are known, the S is given by:

$$S = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$$

Degrees of Freedom:

$$d.f. = n_1 + n_2 - 2$$

Remark: Here it is assumed that the standard deviation (unknown) of the two populations are same.

■ EXAMPLE 12.26

The following data is given:

Sample	Mean	S.D.
A	600 Hrs	63 Hrs
B	700 Hrs	56 Hrs

n for the both sample = 50 each. Discuss if the difference between two means is significant?

Solution:

Null hypothesis: $\mu_1 = \mu_2$
Here $S_1^2 = 63^2$ and $S_2^2 = 56$

$$S = \sqrt{\frac{n_1 S_1^2 + n_2 S_1^2}{n_1 + n_2 - 2}} = \sqrt{\frac{63^2 \times 50 + 56^2 \times 50}{50 + 50 - 2}} = 60.206$$

$$t = \frac{\bar{x_1} - \bar{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{600 - 700}{60.206\sqrt{\frac{1}{50} + \frac{1}{50}}} = 8.3$$

$$d.f. = n_1 + n_2 - 2 = 50 + 50 - 2 = 98$$

Now, we shall find out number of d.f. = (N-1) = $(50 \times 2 - 2)$ = 98. The table value of t, for 98 d.f. at 5% and 1% level of significance are 1.984 and 2.626 respectively. The calculated value of t is several times as large as these limits, hence there is a significant difference between the two means □

■ EXAMPLE 12.27

Two kinds of manure applied to 15 plots of one-acre; other condition remaining the same. The yields (in quintal) are given below:

Manure I	14	20	34	48	32	42	30	44
Manure II	31	18	22	28	40	26	45	

Examine the significance of the difference between the mean yields due to the application kinds of manure

Solution:

Manure I (x_1)	$d_1 = x_1 - 33$	d_1^2	Manure II (x_2)	$d_2 = x_2 - 30$	d_2^2
14	-19	361	31	1	1
20	-13	169	18	-12	144
34	1	1	22	-8	64
48	15	225	28	-2	4
32	-1	1	40	10	100
42	9	81	26	-4	16
30	-3	9	45	15	225
44	11	121			
264	0	968	210	0	554

$$\bar{x_1} = \frac{\Sigma x_1}{n} = \frac{264}{8} = 33; \qquad \bar{x_2} = \frac{\Sigma x_2}{n} = \frac{210}{7} = 30$$

$$S = \sqrt{\frac{\Sigma d_1^2 + \Sigma d_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{968 + 554}{8 + 7 - 2}} = 10.82$$

$$t = \frac{\bar{x_1} - \bar{x_2}}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{33 - 30}{10.82}\sqrt{\frac{8 \times 7}{8 + 7}} = 0.54$$

$$d.f. = n_1 + n_2 - 2 = 8 + 7 - 2 = 13$$

Table value of t for 13 d.f. at 5% level of significant = 2.16. Calculated value of t ¡ Table value of t. Hence, the difference of average yields is not significant □

■ EXAMPLE 12.28

The annual salary of professors in the Government Colleges averages Rs. 25,000 and has standard deviation of Rs. 1,000. In the same State, the salary of doctors averages Rs. 90,000 and has a standard deviation of Rs. 1,500. These data relate to samples of size of 20 for each group chosen at random. Test at 5% level of significance whether there is significance in the mean salaries of two groups Solution

Solution:

$$t = \frac{\bar{x_1} - \bar{x_2}}{S}\sqrt{\frac{N_1 N_2}{N_1 + N_2}}$$

As the standard deviations of both the samples are given, combined S is to be calculated :

$$S = \sqrt{\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2}}$$

$$= \sqrt{\frac{20 \times 1{,}000^2 + 20 \times 1{,}500^2}{20 + 20 - 2}}$$

$$= 1{,}307.87$$

$$t = \frac{\bar{x_1} - \bar{x_2}}{S}\sqrt{\frac{N_1 N_2}{N_1 + N_2}} = \frac{25{,}000 - 20{,}000}{1{,}307.87}\sqrt{\frac{20 \times 20}{20 + 20}} = 12.08$$

$$d.f. = N_1 + N_2 - 2 = 20 + 20 - 2 = 38$$

Table value of t for 38 d.f. at 5% level of significance = 2.02. Computed value of t ¿ Table value of t. The difference the average income of two groups is significant □

12.8.4 Test for Coefficient Correlation in Small Sample

In case of small samples standard error of r is given by

$$S.E._r = \frac{\sqrt{1-r^2}}{\sqrt{n-2}}$$

t-statistic is defined as follows with $(n-2)d.f$

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

Null Hypothesis : The population correlation coefficient is zero.

■ **EXAMPLE 12.29**

It was found that the correlation coefficient between two variables calculated from a sample of size 25 was 0.37. Does this show evidence of having come from a population with zero correlation?

Solution: Null Hypothesis : The population correlation coefficient is zero Given : $n = 25, r = 0.37$

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$
$$= \frac{0.37}{\sqrt{1-0.37^2}} \times \sqrt{25-2} = 1.91$$
$$d.f. = n-2 = 25-2 = 23$$

Table value of t for 23 $d.f.$ at 5% level of significance = 2.069. Calculated value of t ¡ Table value of t. Hence, null hypothesis is accepted. The sample has been taken from a population in which correlation coefficient is zero. □

■ **EXAMPLE 12.30**

Find the least value of r in sample of 18 pairs from a bivariate normal population significant at 5% level

Solution: Given :$n = 18, d.f. = n-2 = 18-2 = 16$

Substitute the given values in

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$
$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{18-2}$$
$$\therefore t = \frac{4r}{\sqrt{1-r^2}}$$

The calculated value of t should be greater than table value of $t = 2.12$ at 5% level of significance for 16 $d.f.$

$$\frac{4r}{\sqrt{1-r^2}} > 2.12$$
$$\Rightarrow 16r^2 > 4.4944(1-r^2)$$
$$\therefore r > \sqrt{\frac{4.4944}{20.4944}} = 0.468$$

□

Problems

12.13 Ten individuals are chosen at random from a population and their height are found to be in inch 63,63,66,57,68,69,70,70,71,71. In the light of these data discuss the suggestion that the mean height in the universe is 66 inch. Given that V=9, P=.947 for t=1.8 and P=.955 for t=1.9.

12.14 The yields of two types *Type 17'* and *Type 51'* of grams in pounds per acre at 6 replications are given below. What comments would you make on the difference in the mean yields? You may assume, that if there be 5 degrees of freedom and p=0.2; t is approximately 1.476

Replication	Yield in lb	
	Type 17	Type 51
1	20.50	24.86
2	24.60	26.39
3	23.06	28.19
4	29.98	30.75
5	30.37	29.97
6	23.83	22.08

Remark : Here the table value of t is given for 5 degrees of freedom. So one may apply paired t-test.

12.15 The following results were obtained from a sample of 10 boxes of biscuits:
Mean weight of contents 490 gm
S.D. of the weight of the contents 9 gm
Test the hypothesis that population mean = 500 gm. Use the following extract of t-table for the purpose

Value of t on 5% level of sign	2.306	2.262	2.228
Value of t on 1% level of sign	3.355	3.250	3.169
Degree of freedom	8	9	10

12.16 The height of six randomly chosen sailors are in inch, 63,65,68,69,71 and 72. The height of 10 randomly chosen soldiers are 61,62,65,66,69,69,70,71,72 and 73. Discuss the light that these data throw on the suggestion that soldiers are on the average taller than sailors.

12.17 The eleven students of a school were given a test in drawing. After it, they were given a months special coaching a second test of equal difficulty was hold at the end of it. Do the following marks give evidence that the students have been benefited by special coaching:

Students	1	2	3	4	5	6	7	8	9	10	11
Marks(I Test)	23	20	19	21	18	20	18	17	23	16	19
Marks(II Test)	24	19	22	18	20	22	20	20	23	20	17

At 5% level of significance for 10 and 11 degrees of freedom, the table values of t are 2.228 and 2.201 respectively.

12.18 A certain medicine given to each of the 9 patients resulted in the following increase of blood pressure. Can it be concluded that the medicine will in general be accompanied by an increase in blood-pressure?
7,3,-1,4,-3,5,6,-4,1.
For 8 degrees of freedom the table value of t at 5% level of significance is 2.306

12.19 The means of two random samples of size 9 and 7 respectively are 196.42 and 198.82 respectively. The sum of the squares of the deviation from the mean are 26.94 and 18.73 respectively. Can the samples be considered to have been drawn from the same normal population?

12.20 A set of 15 observations gives means = 68.57, standard deviation = 2.40, another of 7 observations gives means = 64.14, standard deviation=2.70.

Use the *t* test to find whether the two sets of data were drawn from populations with the same mean, it being assumed that the standard deviation in the two population were equal.

12.21 Sample of size (i) 8 and (ii) 12 are drawn from two batches of certain goods. The coefficient of correlation between two characteristics of the articles are 0.32 and 0.19 respectively. Are these values significant?

12.9 CHI-SQUARE TEST

Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. For example, if, according to Mendel's laws, you expected 10 of 20 offspring from a cross to be male and the actual observed number was 8 males, then you might want to know about the "goodness to fit" between the observed and expected. Were the deviations (differences between observed and expected) the result of chance, or were they due to other factors. How much deviation can occur before you, the investigator, must conclude that something other than chance is at work, causing the observed to differ from the expected. The chi-square test is always testing what null hypothesis, which states that there is no significant difference between the expected and observed result.

12.9.1 Chi-Square Distribution

So for we discussed the sampling distribution of $\bar{x}$, the sample mean. Now let us know about the sampling distribution of sample variance s^2 when samples are randomly drawn from a normal population with mean μ and variance σ^2. Let $x_1, x_2, ...x_n$ be a sample of size n from the normal population, then the sampling distribution of the statistic

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{\sum(x-\bar{x})^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Here χ^2(chi-square) is a statistic which has its distribution as χ^2 with $(n-1)$ degrees of freedom. Another way of defining the χ^2 distribution is as follows:

IF $X_1, X_2, ...X_n$ are independent normal variates with mean μ and variance σ^2 then

$$Z_i = \frac{X_i - \mu}{\sigma}$$

are independent normal variates with mean 0 and variance 1. Then the sum of squares of the variates $Z_1, Z_2, ...Z_n$,i.e.,

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$$

is distributed as χ^2 with n degrees of freedom. Symbolically.

$$\chi^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} (\frac{X_i - \mu}{\sigma})^2 \sim \chi^2_n$$

12.9.2 Some properties of chi-square distribution

1. The distribution of χ^2 lies in the first quadrant.

2. The range of χ^2 distribution is from 0 to ∞.
3. It has a unimodal curve.
4. χ^2-distribution has only one parameter n which is its degrees of freedom (d.f).
5. The shape of the curve greatly vary as n varies.
6. The mean and variance of the χ^2 distribution with n d.f. are:
7. mean = n; variance = $2n$
8. IF χ_1^2 and χ_2^2 are two independent χ^2 variates with n_1 and n_2 degrees of freedom respectively, then their sum $\chi_1^2+\chi_2^2$ will be distributed as chi-square with n_1+n_2 d.f. This addition property of chi-square holds good for any number of χ^2 variates. Thus, if $\chi_1^2, \chi_2^2, ...\chi_k^2$ are k independent χ^2-variates with $n_1, n_2, ...n_k$ d.f. respectively, then their sum $\chi^2 = \sum_{i=1}^{n} \chi_i^2 \sim \chi^2$ with $(n_1+n_2+...+n_k)d.f.$

12.9.3 Test of significance for population variance

Let a random sample of size n be drawn from a normal population with mean μ and variance σ^2, μ and σ^2being unknown. Using the sample information we wish to test the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$,i.e., population variance is σ_0^2. For testing this hypothesis, the test statistic is

$$\chi^2 = \frac{\sum(x-\bar{x})^2}{\sigma_0^2} \sim \chi^2_{(n-1)d.f.}$$

In terms of S^2 or s^2, the statistic in (13) can be written as

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{ns^2}{\sigma_0^2}$$

After computing the calculated value of χ^2 statistic in (13), the acceptance or rejection of H_0 is done according to the following rules -

1. Calculated $\chi^2 \leq \chi^2_{(n-1)}(\alpha)$: Accept H_0 :
2. Calculated $\chi^2 > \chi^2_{(n-1)}(\alpha)$: Reject H_0 :

In case of testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 < \sigma_0^2$ at α level of significance , then if

1. Calculated $\chi^2 \geq \chi^2_{(n-1)}(1-\alpha)$: Accept H_0 :
2. Calculated $\chi^2 < \chi^2_{(n-1)}(1-\alpha)$: Reject H_0 :

Here, $\chi^2_{(n-1)}(\alpha)$ and $\chi^2_{(n-1)}(1-\alpha)$ are critical valves of χ^2- statistics and , for given α and $v = (n-1)$ degrees of freedom, can be seen from table-D in the appendix, For example, for n =10 and $\alpha = 0.05, \chi^2_{(n-1)}(\alpha)(0.05) = 16.9$.

■ **EXAMPLE 12.31**

A random sample of size 25 from a population gives the sample standard derivation to be 9.0. Test the hypothesis that the population standard derivation is 10.5.

Solution: Given that $n = 25, s^2 = \frac{\sum(x-\bar{x})^2}{n} = (9.0)^2 = 81$

To test the null hypothesis $H_0 : \sigma^2 = \sigma_0^2 = (10.5)^2$ we calculate the test statistic

$\chi^2 = \frac{\sum(x-\bar{x})^2}{\sigma_0^2} = \frac{ns^2}{\sigma_0^2} = \frac{25*81}{(10.8)^2} = \frac{2025}{116.64} = 17.36$

also the tabulated value of χ^2 for $(n-1) = 24$ degrees of freedom at 5 % level is $\chi^2_{24}(0.05) = 36.42$.

Now since the calculated value of $\chi^2 = 17.36 < \chi^2_{24}(0.05) = 36.42, H_0$ is accepted and we conclude that the population standard deviation may be 10.5. □

■ **EXAMPLE 12.32**

Weights in kg. of 10 students are given below :

$$38, 40, 45, 53, 47, 43, 55, 48, 52, 49$$

Can we say that variance of the distribution of weights of all students from which the above sample was taken is equal to 20 square kg.

Solution: Have, we wish to test $H_0 : \sigma^2 = \sigma_0^2 = 20$. To test H_0, the test statistic is
$\chi^2 = \frac{\sum(x-\bar{x})^2}{\sigma_0^2}$
Here $\bar{x} = \frac{\sum x}{n} = \frac{470}{10} = 47$
Putting the values in (i),
$\chi^2 = \frac{280}{20} = 14.$
Also $\chi_9^2(0.05) = 16.92.$

Thus, the calculated value of χ^2 is less than the tabulated value. the null hypothesis holds true and we conclude that population variance may be 20 square k.g. □

12.9.4 Testing the Goodness of Fit

So far we have discussed the testing of hypothesis about the population parameters like μ, σ^2 and p. Now we consider a test to determine if a population has a specified theoretical distribution. In other words, here our problem is to test the hypothesis of how closely the observed distribution approximates a particular theoretical distribution. As a example, let us consider a die tossing experiment in which the die is tossed 900 times. Now, under the assumption that the die is fair, the theoretical frequency for each face will be 15 . On the other hand, observed frequencies for each face in actual tossing of a die 90 times may differ from theoretical frequencies as shown in the following table :

Observed and theoretical frequencies of 900 tosses of a die

Faces	1	2	3	4	5	6	Total
Theoretical or Expected frequencies	15	15	15	15	15	15	90
Observed Frequencies	18	12	11	19	16	14	90

By comparing the observed frequencies with the corresponding expected frequencies we wish to test whether the differences are likely to occur due to fluctuation of sampling and the die is unbiased, or the die is biased and the differences are real .

For explaining the goodness of fit test, let us consider a population which may be partitioned or classified into k classes, and let p_i be the probability that an observations belongs to the ith class $(i = 1, 2, ...k)$ with $\sum_{i=1}^{k} p_i = 1$, Further let a random sample of size n be drown from the population. Then suppose o_i is the number of sample observations belonging to ith class and obviously, $\sum_{i=1}^{k} o_i = n$. Also let e_i be the expected frequency in ith class as computed under the null hypothesis . Thus , $e_i = E(o_i) = np_i, i = 1, ...k$. Sum of the expected frequencies too will be n , i.e., $\sum_{i=1}^{k} e_i = n$.Then , data can be put as in the following table :

Class	1	2	3	...i	...k	Total
Observed freq. o_i	o_1	o_2	o_3	$...o_i$	$...k$	n
p_i	p_1	p_2	p_3	$\cdots p_i$	$\cdots p_k$	n
$e_i = np_i$	e_1	e_2	e_3	$...e_i$	$...n_k$	n

Now, a goodness of fit test between observed and expected frequency is based on the statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where the sampling distribution of the statistic χ^2 in (12) is $(k-1)$ degrees of freedom.

From χ^2 value in (14), we see that if the observed frequencies (o_i) are close to the expected frequencies (c_i), the χ^2 value will be small , which indicates that the fit is good. On the other hand , if $o_i's$ and $e_i's$ differ considerably, the χ^2 will be large and fit is not good. thus , acceptance of H_0 : means the fit is good and its rejection means that the fit is not good.

The rejection and acceptance of H_0 at α level of significance depends on the critical value of χ^2- statistics obtained from table-D given in the appendix. this critical value is seen for α level and $v = (k-1)d.f.$ and is denoted as $\chi^2_{(k-1)}(\alpha)$.Now, if

(i) Calculated $\chi^2 \leq \chi^2_{(k-1)}(\alpha)$; Accept H_0.

(ii) Calculated $\chi^2 > \chi^2_{(k-1)}(\alpha)$; Reject H_0

12.9.4.1 Important Remarks

1. In the goodness of fit test observed frequencies in the k class are given while expected frequencies are obtained by assuming the null hypothesis to be true or that the fit is good.
2. Before calculating the value of χ^2statistic if we observe that the expected frequencies in any class is less than 5, then such frequencies are pooled or combined with adjacent classes. Consequently , the corresponding observed frequencies too are pooled or combined with the adjacent classes and in the process, the number of classes is also reduced. for example. Let us consider the following situation in which we have six classes.

Classes	1	2	3	4	5	6	Total
o_i	3	1	10	15	6	3	38
e_i	1	4	9	18	4	2	38

See that the expected frequencies in first , second, fifth and sixth classes are less than 5, so we combine these frequencies as shown in the adjacent table.

In this way, after combining the frequencies , the number of classes reduces to 3 only.

Pooled frequencies

Classes	o_i	e_i
1	3	1
2	1	4
3	10	9
4	15	18
5	6	4
6	3	2
Total	38	38

■ EXAMPLE 12.33

In an experiment on pea breeding. Mendal obtained the following frequencies of seeds: 315 round and yellow; 101 wrinkled and yellow; 108 round and green; 32 wrinkled and green. Total 556. Theory predicts that the frequencies would be in the proportion 9 : 3 : 3 : 1. Does the experimental results support the theory?

Solution: H_o : Let the Mendalin assumption be true , i.e., the frequencies of peas in the four classes are in the ratio 9 : 3 : 3 : 1.

Thus, the expected frequency of round and yellow seeds
$E_1 = \frac{9}{16} \times 556 = 313$
The expected frequency of wrinkled and yellow seeds
$E_2 = \frac{3}{16} \times 556 = 104$
The expected frequency of round and green seeds
$E_3 = \frac{3}{16} \times 556 = 104$
the expected frequency of wrinkled and green seeds
$E_4 = \frac{1}{16} \times 556 = 35$
Calculation of χ^2

S.No.	Observed frequency (O_i)	Expected frequency E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i-E_i)^2}{E_i}$
1	315	313	2	4	0.0128
2	101	104	-3	9	0.0865
3	108	104	+4	16	0.1538
4	32	35	-3	9	0.2812
Total	556	556			0.5103

$$\therefore \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 0.5103$$

Also

$$\chi^2_3(0.05) = 7.815$$

The calculated value of χ^2_3 is much less than the table value. Thus, we accept H_0 and conclude that there seems a correspondence between theory and experiment. □

■ EXAMPLE 12.34

The following table gives the no. of aircraft accidents that occurred during the various days of the week. Find whether the accidents are uniformly distributed over the week. $(\chi^2_{6,0.05} = 12.59)$

Days	Sun	Mon	Tue	Wed	Thu	Fri	Sat
No. of Accidents	14	15	8	20	11	9	14

Solution: Here we setup the null hypothesis that the accidents are uniformly distributed over the week.

According H_0, the expected frequency of accidents for each day will be

$$= \frac{\text{Total Number of accidents}}{6} = \frac{91}{7} = 13$$

The calculation of χ^2 is shown in the table.

S.No.	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i-E_i)^2}{E_i}$
1	14	13	1	1	0.0769
2	15	13	2	4	0.3077
3	8	13	-5	25	1.9231
4	20	13	7	49	3.7692
5	11	13	-2	4	0.3077
6	9	13	-4	16	1.2308
7	14	13	1	1	0.0769
Total	91	13			$\sum \frac{(O_i-E_i)^2}{E_i} = 7.6923$

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 7.6923.$$

Also

$$\chi^2_5(0.05) = 12.59$$

since calculated χ^2 is less than the tabulated value, so we accept H_0 and conclude that accidents are uniformly distributed over the week. □

■ EXAMPLE 12.35

A die is thrown 90 times and the number of faces are as indicated below.

Faces	1	2	3	4	5	6
Frequency	18	14	13	15	14	16

Solution: We set up the hypothesis that the die is fair . Thus, with this hypothesis the expected frequency for each face will be :

$$= \frac{90}{6} = 15$$

Thus,

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(18-15)^2}{15} + \frac{(14-15)^2}{15} + \frac{(13-15)^2}{15} + \frac{(15-15)^2}{15} + \frac{(14-15)^2}{15} + \frac{(16-15)^2}{15} \\
&= \frac{(9+1+4+0+1+1)}{15} = \frac{16}{15} = 1.07
\end{aligned}$$

Also

$$\chi^2_5(0.05) = 11.07$$

As the calculated χ^2 is much less than the table at 5% level of significance , so H_0 is accepted and conclude that the die is fair. □

■ EXAMPLE 12.36

100 students of a management institute obtained the following grades in statistics paper :

Grade	A	B	C	D	E	Total
Frequency	15	17	30	22	16	100

Using χ^2test , examine the hypothesis that the distribution of grades is uniform.

Solution: Here we set up the null hypothesis that the grades are uniformly distributed among the students. According to H_0, the expected frequency of grades will be

$$= \frac{\text{Total number of students}}{5} = \frac{100}{5} = 20$$

The Calculation of χ^2 is shown in the table :

S.No.	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
1	15	20	-5	25	1.25
2	17	20	-3	9	0.45
3	30	20	10	100	5.00
4	22	20	2	4	0.20
5	16	20	-4	16	0.80
Total	100				$\sum \frac{(O_i-E_i)^2}{E_i} = 7.7$

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 7.7$$

Also

$$\chi_4^2(0.05) = 9.49$$

Such calculated χ^2 is less than the tabulated value, so we accept H_0 and conclude that distribution of grades is uniform. □

■ EXAMPLE 12.37

Five dice were thrown 96 times and the number 4,5, or 6 occurred as given below:

No of throwing (4,5 or 6)	5	4	3	2	1	0
frequency	7	19	35	24	8	3

Calculate χ^2. Find the probability of getting this result.

Solution: Here the theoretical frequencies are given by binomial distribution

$$N^n C_r p^r (1-p)^{n-r}$$

which are 3,15,30,30,15,3. Since no frequency should be less than 5 for χ^2 test, so we change our problem as

No of throwing (4,5 or 6)	5 or 4	3	2	1 or 0
frequency (O_i)	7+19=26	35	24	8+3=11
frequency (E_i)	3+15=18	30	30	15+3=18

$$\begin{aligned}\chi^2 &= \frac{\Sigma(O_i - E_i)^2}{E_i} \\ &= \frac{(26-18)^2}{18} + \frac{(35-30)^2}{30} + \frac{(24-30)^2}{30} + \frac{(11-18)^2}{18} \\ &= 8.31\end{aligned}$$

To find the probability of this result we have, the degree of freedom=4-1=3. Now for $\chi^2 = 7.815$ and degree of freedom 3, we have P=0.05 and for $\chi^2 = 9.837$ and degree of freedom 3, the probability is 0.02

When difference in χ^2 is 2.022, the difference in P is -0.03. and difference of 8.31 fro, 7.815 is 0.495. Therefore difference in P corresponding to 0.495

$$\frac{-0.003 \times 0.495}{2.022} = -0.007$$

$$\therefore P(\text{ for } \chi^2 = 8.31) = 0.05 - 0.007 = 0.043$$

□

12.9.5 Contingency Table

Suppose N observations in a sample are to be classified according to two attributes A and B . Attribute A has r mutually exclusive categories say $A_1, A_2, ...A_r$ and the attribute B has s categories namely $B_1, B_2, \ldots, B_s$. then the sample observations may be classified as shown:

A\B	B_1	B_2	...	B_j	...	B_s	Total
A_1	O_{11}	O_{12}	...	O_{1j}	...	O_{1s}	(A_1)
A_2	O_{21}	O_{22}	...	O_{2j}	...	O_{2s}	(A_2)
:	:	:	:	:	:	:	:
A_i	O_{i1}	O_{i2}	...	O_{ij}	...	O_{is}	(A_i)
:	:	:	:	:	:	:	:
A_r	O_{r1}	O_{r2}	...	O_{rj}	...	O_{rs}	(A_r)
Total	(B_1)	(B_2)	...	(B_j)	...	(B_s)	N

The above two-way table having r rows and s columns is called a contingency table of order (r * s) = In This table :

1. A_i denotes the ith category of the attribute $A, (i = 1, 2, ...r)$
2. B_j denotes the jth category of the attribute $B, (j = 1, 2, ...s)$
3. (A_i) denotes the frequency of the attribute A_i.
4. (B_j) denotes the frequency of the attribute $B_{j.}$
5. $\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{s}(B_j) = N$, the total number of observations.
6. (O_{ij}) is the observed frequency in (i, j) cell $[(i = 1, 2, \ldots, r), (j = 1, 2, \ldots, s)]$.

Important Remark. If two attributes A and B are independent, then the expected frequency in (i,j) the cell in a contingency table, i.e., E_{ij}, is

$$E_{ij} = E(O_{ij}) = \frac{(A_i)(B_j)}{n}, \ [i = 1, 2, \ldots, r), (j = 1, 2, \ldots, s)]$$

12.9.6 Testing the independence of two attribute in a contingency table

Suppose we are given a contingency table of order $(r \times s)$ in which N sample observations have been classified. Let O_{ij}be the observed frequency in (i, j)th cell. To test the null hypothesis that the two attributes are independent, we use χ^2-test. First we set up the null hypothesis:

H_0 : Two attributes A and B are independent.

H_1 : Two attributes are dependent or associated.

To test H_0, we calculated the test statistic:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

where, the statistic χ^2in (15) χ^2 distribution with $(r-1)(s-1)$ degrees of freedom. the expected frequency corresponding to (i, j)th observed frequency, i.e,E_{ij} is obtained by assuming H_0 to be true.

Therefore,

$$\begin{aligned} E_{ij} = E(O_{ij}) &= \frac{A_i \times B_j}{N} \\ &= \frac{\text{Sum of } i\text{th row} \times \text{Sum of } j\text{th column}}{\text{Sample size}} \end{aligned}$$

$$(i = 1,2,\ldots,r \text{ and } = 1,2,\ldots,,s)$$

After getting E_{ij}'s or expected frequencies in the cells ,χ^2value is calculated by using the formula in (15.) The decision about the acceptance or rejection of $H_0 at$of significance is taken as under:

(i) If calculated $\chi^2 < \chi^2_{(r-1)(s-1)}(\alpha)$; AcceptH_0

(ii) If calculated $\chi^2 > \chi^2_{(r-1)(s-1)}(\alpha)$; Accept H_0.

$\chi^2_{(r-1)(s-1)}(\alpha)$ is the critical value of χ^2-statistic for degree of freedom $v = (r-1)(s-1)$.and for αlevel of significance The value of $\chi^2_v(\alpha)$ are given in Table-D of the Appendix.

Remark.

1. Acceptance of H_0means that the two attributes are independent and there is no association between then.
2. Rejection of H_0leads to the conclusion that $H_1$is true and data support the hypothesis that there is some relationship between the two variables.

EXAMPLE 12.38

The following table shows the classification of 4000 workers in a factory, according to the disciplinary action taken by the management of the factory and their promotional experience.

Disciplinary Action	Promotional Experience		Total
	Promoted	Not promoted	
Non-offenders	300	900	1200
offenders	100	2700	2800
Total	400	3600	4000

Use x^2test to find if there is any evidence to support that there was any association between disciplinary action and the promotional experience of the workers in this factory at 5% level of sesnificance.

Degree of freedom	1	2	3	4	5
$x^2 - 5\%$	3.84	5.99	7.82	9.49	11.07

Solution: We form the hypothesis that disciplinary action and promotional experience of the workers are independent. Assuming to be the hypothesis true, we find the expected frequencies for each cell.

Disciplinary Action	Promotional Experience		Total
	Promoted	Not promoted	
Non-offenders	$\frac{400\times 1200}{4000}=120$	$\frac{1200\times 3600}{4000}=1080$	1200
Offenders	$\frac{400\times 2800}{4000}=280$	$\frac{2800\times 3600}{4000}=2520$	2800
Total	400	3600	4000

$$\chi^2=\frac{(300-120^2)}{120}+\frac{(900-1080)^2}{1080}+\frac{(100-280)^2}{280}+\frac{(2700-252)^2}{2520}$$
$$=\frac{(180)^2}{120}+\frac{(180)^2}{1080}+\frac{(180)^2}{280}+\frac{(180)^2}{2520}=428.57$$

Also table value of χ^2at 1 degree of freedom for 5% level of significance is 3.841. Since calculated value of $x^2$is much greater than table value of χ^2, we reject $H_0$and conclude that disciplinary action and promotional experience of the workers are dependent. □

■ EXAMPLE 12.39

The following table given the number of accounting clerks committing errors and not committing errors among trained and untrained clerks working in an organization:

	Number of clerks committing errors	Number of clerks not committing errors	Total
Trained	70	530	600
Untrained	155	745	900
Total	225	1275	1500

Solution: Here we set up the null hypothesis that training and committing error are independent. Using the null hypothesis the expected frequencies in different cells are as n obtained in the following table:

	Number of clerks committing errors	Number of clerks not committing error	Total
Trained	$\frac{225\times 600}{1500}=90$	$\frac{1275\times 600}{1500}=510$	600
Untrained	$\frac{1275\times 900}{1500}=135$	$\frac{1275\times 900}{1500}=765$	900
Total	225	1275	1500

T_0test $H_{0,}$the test statistic is :

$\therefore$

$$\chi^2=\sum_i\sum_j\frac{(O_{ij}-E_{ij})^2}{E_{ij}}=\frac{(70-90)^2}{90}+\frac{(530-510)^2}{510}+\frac{(155-135)^2}{135}+\frac{(745-765)^2}{765}$$
$$=\frac{(20)^2}{90}+\frac{(20)^2}{510}+\frac{(20)^2}{135}+\frac{(20)^2}{765}=8.71$$

Here $c=2$ and $r=2 \therefore d.f.=(2-1)(2-1)=1.$

Thus, $x^2 = 8.17 > x^2(.05) = 3.84$, we reject, H_0 at 5% level of signification and conclude that training is effective in preventing errors. □

Problems

12.22 In the course of anti-malaria works, quinine was administered to 606 males out of a total population of 3540. The incidence of malaria fever is shown below:

	Infection	No infection
Administering quinine	19	587
Not administering quinine	193	2741

12.23 The following table shows price increase decrease in market, where credit squeeze is in operation and where it is not in operation:

	Increase	Decrease
In operation	862	10
In Not operation	582	18

Find whether the credit squeeze has been effective in checking price increase.

12.24 Two treatments A and B were tried to a certain type of plant disease. 100 plants treated A were examined and 20 were found infected. 100 plants treated B were examined and 5 were found infected. Is treatment B is superior than A? 5% value of χ^2 for one degree of freedom is 3.841.

12.25 A sample analysis of examination results of 500 students was made. It was found that out of 500 students, 220 students had failed, 170 had secured third class and 90 were placed in second, other got first class. Are these figures commensurate with the general examination results which is the ratio of 4:3:2:1 for the various categories respectively.

CHAPTER 13

ANALYSIS OF VARIANCE

We have studied the test of significance of sample mean from the mean of the universe or the test of difference of two sample means. For this we used the standard error of mean or the standard error of difference of the two means, using z-test or t-test. But in this period of scientific experiments to test the significance between more than two means, is not possible by these methods.

Suppose we want to study the effects of four types of fertilizers, say A, B, C and D on the yield of sugar cane. We take five plots for each fertilizer. In this way, the use of 4 fertilizers is done on 20 plots. We can find the arithmetic means of the yields of 5 plots for each fertilizer separately. But the test of significance of the difference of these means is not possible with t-test. However, one way using t-test is that we make 6 pairs of two fertilizers AB, AC, AD, BC, BD and CD and then test their difference. Conclusion can also be drawn separately. There arise two difficulties:

1. First, the work of computation will increase and

2. Second, Only the pairs are tested out of the four fertilizers. We can not find whether the difference is significant taking them together.

In such situation a method of test of significance to avoid these two difficulties, is needed and the desired objective test of significance between the means of more than two samples is fulfilled. Here test of significance means, to test the hypothesis whether the means of several samples have significant difference or not. In other words, to test whether the samples can be considered as having been drawn from the same parent population or not and the difference among the means is due to sampling of fluctuations or due to some other cause. To test the difference among several sample means we use a statistical technique known as *Analysis of Variance*.

13.1 MEANING OF ANALYSIS OF VARIANCE

Analysis of variance basically is an arithmetical method by which we split up the total variability into component variations ascribable to different sources of causes. In the words of Yule and Kendall,

> The analysis of variance is essentially a procedure for testing the difference between different groups of data for homogeneity.

While defining analysis of variance Sir, Ronal A. Fisher wrote,

> ... The separation of the variance ascribable to one group of causes from the variance ascribable to other groups.

In simple words, Analysis of Variance is a statistical technique, with help of which total variation, is partitioned into variation caused by each set of independent factors and homogeneity of several means is tested.

13.1.1 Components of total Variability

In general (say one way classification) total variability is partitioned into two parts that is :

Total Variability = Variability between samples + Variability within samples.
Or
Total Variation = Variation between samples + Variation within samples.

13.1.2 Assumptions of Analysis of Variance

The analysis of variance is based on certain assumptions as given below :

1. Normality of the Distribution : The population for each sample must be normally distributed with mean μ and unknown variance σ^2.
2. Independence of Samples : All the sample observations must be selected randomly. The total variation of the various sources of variation should be additive.
3. Additivity : The total variation of the various sources of variation should be additive.
4. Equal variances (but unknown) : The populations from which the n samples say are drawn have means μ_1, μ_2, ..., μ_n and unknown variance $\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = \sigma^2$.
5. The error components are independent and have mean 0 and variance σ^2.

The tests of significance performed in the analysis of variance are meaningful under its assumptions.

13.2 ONE-WAY CLASSIFICATION

When observations are classified into groups or samples on the basis of single criterion, then it is called One-way classification. For examples, The yield of sugar cane of 20 plots, classified in plots on the basis of four types of fertilizer, the marks obtained by students of different colleges, etc.

13.2.1 Techniques of One-way Analysis of Variance

1. In One-way analysis of variance there are k samples, one from each of k normal populations with common variance σ^2 and means $\mu_1, \mu_2, ...\mu_n$. The number of observations n_i in samples may be equal or unequal i.e.

$$n_1 + n_2 + ... + n_k = N$$

2. Linear Model

$$
\begin{aligned}
x_{ij} &= \mu + \alpha_i + e_{ij} \\
\text{Where } x_{ij} &= \text{ observations } i = 1, 2, \ldots.k,\ j = n_i \\
\mu &= \text{ The general mean} \\
\alpha_i &= \text{ Effect of } i^{th} \text{ factor } = \mu_i - \mu \\
e_{ij} &= \text{ Effect of error or random term.}
\end{aligned}
$$

3. Null Hypothesis (H_0) and Alternative Hypothesis (H_1):

   H_0 : The means of the populations are equal i.e.

$$\mu_1 = \mu_2 = \ldots., = \mu_k$$

   H_1: At least two of the means are not equal.

4. Computations:

   (i) Calculate sum of observations in each sample and of all observations.

   Sum of sample observations : $\sum x_1, \sum x_2, \ldots\ldots \sum x_k$.

   Sum of the squares of the sample observations : $\sum x_1^2, \sum x_2^2, \ldots\ldots \sum x_k^2$

   (ii) Calculate correction factor CF = $\frac{T^2}{N}$

   Where T = Square of the sum of all the observations = $\sum x$

   N = Total number of observations

   (iii) Calculate sample means $\bar{X_1}, \bar{X_2}, \ldots.., \bar{X_k}$ and their common mean $\bar{\bar{X}}$. where $\bar{X_k}$ = $\frac{\Sigma x_k}{n_k}$, $\bar{\bar{X}}$ = $\frac{\Sigma x}{N}$

   (iv) Calculate total sum of squares (TSS) by the formula

$$
\begin{aligned}
TSS &= \Sigma(x - \bar{\bar{X}})^2 \\
&= (\Sigma x_1^2 + \Sigma x_2^2 + \ldots.. + \Sigma k_n^2) - \frac{T^2}{N} \\
&= \Sigma x^2 - \frac{(\Sigma x)^2}{N}
\end{aligned}
$$

   (v) Sum of squares between (SSB) samples by the formula

$$
\begin{aligned}
SSB &= n_1(\bar{X_1} - \bar{\bar{X}})^2 + n_2(\bar{X_2} - \bar{\bar{X}})^2 + \ldots + n_k(\bar{X_k} - \bar{\bar{X}})^2 \\
&= \left[\frac{(\Sigma x_1)^2}{n_1} + \ldots + \frac{\Sigma x_k)^2}{n_k}\right] - \frac{T^2}{N}
\end{aligned}
$$

   (vi) Calculate sum of squares within samples by the formula

$$
\begin{aligned}
SSW &= (x_1 - \bar{X_1})^2 + (x_2 - \bar{X_2})^2 + \ldots + (x_k - \bar{X_k})^2 \\
&= \Sigma(x_k - \bar{X_k})^2
\end{aligned}
$$

   Sum of squares may also be computed as

$$SSW = TSS - SSB$$

   Sum of squares within samples is also called Error sum of squares.

(vii) Calculate mean sum of squares :

$$MSSB = \text{Mean sum of squares between samples}$$
$$= \frac{SSB}{k-1} = \frac{\text{Sum of squares between samples}}{\text{Degrees of freedom}}$$
$$MSSW = \text{Mean sum of squares within samples}$$
$$= \frac{SSW}{N-k} = \frac{\text{Sum of squares within samples}}{\text{Degrees of freedom}}$$

Total number of degrees of freedom $= N-1$

$$\text{where } N-1 = (k-1)+(N-k)$$

(viii) Obtain the variance ratio F :

Variance ratio,
$$F = \frac{MSSB}{MSSW} = \frac{\text{Variance between samples}}{\text{Variance within samples}}$$

Remark : In general, $MSSB$ is greater than $MSSW$, so $MSSB$ is taken in the numerator.

(ix) Interpretation of $F - Ratio$:

Compare the calculated value of F with tabulated value.

Let

$F_e =$ Calculated value of F

$F_t = F_{05}(v_1, v_2)$

$=$ Tabulated value of F at 5 level of significance with degrees of freedom v_1 and v_2

Here

v_1 = degree of freedom for numerator

v_2 = degree of freedom for denominator

If $F_c > F_t$, i.e. calculated value of F exceeds the tabulated value of F, we say the difference among sample means is significant and conclude that all the population means are not equal i.e., reject the null hypothesis.

If $F_c < F_t$, the difference among sample means sample means is not significant i.e., accept the null hypothesis.

(x) The results of calculations can be presented in tabular form. This table is called analysis of variance table.

13.3 ANALYSIS OF VARIANCE TABLE

Source of Variation	Sum of Squares SS	Degree of Freedom d.f.	Mean sum of squares MSS	F_c	F_t
Between Samples	$SSB = \Sigma n_k(\bar{x_k} - \bar{x})^2$	$k-1$	$\frac{SSB}{d.f} = MSSB$	$\frac{MSSB}{MSSW}$ SSB	
Within samples	$SSW = \Sigma(x_k - \bar{\bar{X}})^2$	$M-K$	$\frac{SSW}{d.f.} = MSSW$	SSW	
Total	$\Sigma(x - \bar{\bar{X}})^2$	$N-1$			

Remark : There are three methods, to calculate mean and variances :

(i) Direct Method

(ii) Indirect Method
(iii) Step deviation Method or Coding Method (change of origin or / and scale)

EXAMPLE 13.1

Direct Method: Apply the technique of analysis of variance to the following data, relating to yields of 4 varieties of wheat in 3 blocks:

Varieties	Blocks 1	2	3
I	10	9	8
II	7	8	6
III	8	5	5
IV	5	8	5

Test whether the varieties are significantly different with regard to yield Ignoring variation between blocks.
Solution:

Table 13.1. Computation of Arithmetic Mean

Blocks	Varieties I	II	III	IV
1	10	7	8	5
2	9	8	5	8
3	8	6	5	5
Total	27	21	18	18
Mean($\bar{x}$)	9	7	6	6

Mean of all means

$$(\bar{\bar{x}}) = \frac{9+7+6+6}{4} = \frac{28}{4} = 7$$

Variance between Samples. Sum of the squares of the deviations

$$= n_1(\bar{x_1} - \bar{\bar{x}})^2 + n_2(\bar{x_2} - \bar{\bar{x}})^2 + n_3(\bar{x_3} - \bar{\bar{x}})^2 + n_4(\bar{x_4} - \bar{\bar{x}})^2$$
$$= 3(9-7)^2 + 3(7-7)^2 + 3(6-7)^2 + 3(6-7)^2 = 18$$

Degrees of Freedom $df_1 = v_1 = k-1 = 4-1 = 3$
Mean sum of squares between samples (i.e. blocks)

$$= \frac{\Sigma[nk(\bar{x_k} - \bar{\bar{x}})^2]}{(k-1)} = \frac{18}{3} = 6$$

Variance within Samples. Sum of the squares of the deviations

$$= \Sigma(x_1 - \bar{x_1})^2 + \Sigma(x_2 - \bar{x_2})^2 + \Sigma(x_3 - \bar{x_3})^2 + \Sigma(x_4 - \bar{x_4})^2$$
$$= 2+2+6+6 = 16$$

Degree of Freedom $df_2 = v_2 = N - k = 12 - 4 = 8$

Mean sum of squares within samples (i.e. blocks) $= \frac{\Sigma}{(}x-\bar{x_k})^2N-k = \frac{16}{8} = 2$

$$\text{F-Ratio } = \frac{MSSB}{MSSW} = \frac{6}{2} = 3$$

Analysis of Variance.

source of variation	S.S.	$d.f.\ (k-1)$	$M.S.S.\left(\frac{S.S.}{d.f.}\right)$	F-Ratio
Between Samples	18	3	6	$F = \frac{MSSB}{MSSW} = \frac{6}{2} = 3$
within samples	16	$(N-k) = 8$	2	

Conclusion. F or $v_1 = 3$, $v_2 = 8$, $F_t = 4.07$ at 5 level of significance. Since $F_c < F_t$, hence the four varieties of wheat are not significantly different with regard to yield i.e., This difference is due to fluctuations of sampling.

Varieties.

	I	II	III	IV
	$(10-7)^2 = 9$	$(7-7)^2 = 0$	$(8-7)^2 = 1$	$(5-7)^2 = 4$
	$(9-7)^2 = 4$	$(8-7)^2 = 1$	$(5-7)^2 = 4$	$(8-7)^2 = 1$
	$(8-7)^2 = 1$	$(6-7)^=1$	$(5-7)^2 = 4$	$(5-7)^2 = 4$
Total	14	2	9	9

Total sum of Squares = $14+2+9+9$ = 34
Degree of Freedom $= N-1 = 12-1 = 11$
Total Variation = $frac\Sigma(x-\bar{x})^2N-1$ = $\frac{34}{11}$ = 3.09

■ EXAMPLE 13.2

Indirect Method: The following table gives the result of experience of four varieties of wheat grown in 24 plots:

Plot Yield			
A	B	C	D
4	9	9	12
2	5	20	5
5	8	18	15
8	12	14	9
6	14	18	16
12	11	10	17

Is there any significant difference in production of these varieties?

Solution: Null Hypothesis (H_0) : There is no difference between the yields of four varieties of wheat.

Sum of the Squares of the Sample Values.

A		B		C		D	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	x_4	x_4^2
4	16	9	81	9	81	12	144
2	4	5	25	20	400	5	25
5	25	8	64	18	324	15	225
8	64	12	144	14	196	9	81
6	36	14	196	18	324	16	256
		12	144	11	121	10	100
						17	289
25	145	60	654	90	1446	84	1120

$$T = \Sigma x_1 + \Sigma x_2 + \Sigma x_3 + \Sigma x_4 = 25 + 60 + 90 + 84 = 259$$

Correction Factor.

$$C.F. = \frac{T^2}{N} = \frac{(259)^2}{24} = \frac{67,081}{24} = 2,795.04$$

Total Sum of Squares.

$$TSS = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2 - \frac{T^2}{N}$$

or

$$TSS = 145 + 654 + 1,446 + 1,120 - 2,795.04 = 569.96$$

Between Samples S.S. = SSB.

$$SSB = \left[\frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \frac{(\Sigma x_3)^2}{n_3} + \frac{(\Sigma x_4)^2}{n_4}\right] - \frac{T^2}{N}$$
$$= \frac{(25)^2}{5} + \frac{(60)^2}{6} + \frac{(90)^2}{6} + \frac{(84)^2}{7} - 2,795.04 = 287.96$$

Within Samples S.S. = SSW.

$$SSW = TSS - SSB = 569.96 - 287.96 = 282$$

Table 13.2. Analysis of Variance Table

Variance Source	S.S.	d.f	M.S.	F
Between samples	287.96	(4-1)=3	95.99	$\frac{95.99}{14.10}$
Within Samples	282	$(24-4)$=20	14.10	=6.807

Conclusion : $F_{cal} = 6.000$. F_{table} for $v_1 = 3$ and $v_2 = 20$ degrees of freedom at 5% level of significance $= 3.10$. Since $F_{cal} > F_{table}$, the null hypothesis is rejected. That is the difference among the average yield of four kinds of wheat is significant.

EXAMPLE 13.3

Coding Method: Set up a table of analysis of variance for the following data:

	Variety			
Plots	A	B	C	D
1	200	230	250	300
2	190	270	300	270
3	240	150	145	180

Test whether the varieties are different.

Solution: In this method we can change the origin or scale or both as the case may be and proceed as usual. We know that F is independent of the change of origin and scale. So there is no need to adjust the value of $F_{cal.}$

Coded Data			
A	B	C	D
0	+30	+50	+100
-10	+70	+100	+70
+40	-50	-55	-20
$\Sigma x_1 = 30$	$\Sigma x_2 = 50$	$\Sigma x_3 = 95$	$\Sigma x_4 = 150$

Squares			
A	B	C	D
0+900+2	500+10	000	
100	4,900	10,000	4,900
1,600	2,500	3,025	400
$\Sigma x_1^2 = 1,700$	$\Sigma x_2^2 = 8,300$	$\Sigma x_3^2 = 15,525$	$\Sigma x_4^2 = 15,300$

$$
\begin{aligned}
T &= \Sigma x_1 + \Sigma x_2 + \Sigma x_3 + \Sigma x_4 = 30 + 50 + 95 + 150 = 325 \\
C.F. &= \frac{T^2}{N} = \frac{(325)^2}{12} = \frac{1,05,625}{12} = 8,802.08 \\
SST &= \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2 - \frac{T^2}{N} \\
&= 1,700 + 8,300 + 15,525 + 15,300 - 8,802.08 = 32,022.92 \\
SSB &= \frac{(\Sigma x_1^2)}{n_1} + \frac{(\Sigma x_2^2)}{n_2} + \frac{(\Sigma x_3^2)}{n_e} + \frac{(\Sigma x_4^2)}{n_4} - \frac{T^2}{N} \\
&= \frac{(30^2)}{3} + \frac{(50^2)}{3} + \frac{(95^2)}{3} + \frac{(150^2)}{3} - 8,802.08 = 2,839.52 \\
SSW &= TSS - SSB = 32,022.98 - 2,839.52 = 29,183.40
\end{aligned}
$$

Degree of freedom. Between Samples $v_1 = k - 1 = 4 - 1 = 3$, Within Samples $v_2 = N - k = 12 - 4 = 8$, Total $= N - 1 = 12 - 1 = 11$

Analysis of Variance Table.

Source	S.S.	d.f.	M.S.	F-Ratio
Within Samples	29,183.40	8	3,647.93	$F=\frac{3,647.93}{946.51} < 1$
Between Samples	2,839.52	3	946.51	
Total	32,022.92	11		

Interpretation. The calculated value of F is less than 1. Table value of F_t for $v_1 = 3$ and $v_2 = 8$ at 5 level of significance is 4.07. In such cases one may say that the null hypothesis is true, but the fact is that within sum of squares is too much which hides the actual difference between samples if any.

EXAMPLE 13.4

Analysis of variance applied to grouped data i.e., two frequency distributions: Two samples of 20 and 25 students score the following marks in a test carrying maximum marks 15:

Marks	10	11	12	13	14	15	Total
f_1	3	7	6	3	1	0	20
f_2	1	3	5	8	7	1	25

Does the average level of knowledge of the students differ?

Solution: .

x	Marks Ax^2	x^2 f_1	I Sample f_1x	II Sample f_1x^2	f_2	f_2x	f_2x^2
10	100	3	30	300	1	10	100
11	121	7	77	847	3	33	363
12	144	6	72	864	5	60	720
13	169	3	39	507	8	104	1,352
14	196	1	14	196	7	98	1,372
15	225	0	0	0	1	15	225
	Σ	20	232	2,714	25	320	4,132

$$T = \Sigma f_1x + \Sigma f_2x = 232 + 320 = 552$$

$$C.F. = \frac{T^2}{N} = \frac{T^2}{f_1 + f_2} = \frac{3,04,704}{45} = 6,771.2$$

$$TSS = \Sigma f_1x^2 + \Sigma f_2x^2 - C.F.$$

$$= 2,174 + 4,132 - 6,771.2 = 74.8$$

$$SSC = \frac{(\Sigma f_1x)^2}{n_1} + \frac{(\Sigma f_2x)^2}{n_2} - C.F.$$

$$= \frac{(232)^2}{20} + \frac{(320)^2}{25} - 6,771.2 = 16$$

$$SSR = TSS - SSC = 74.8 - 16 = 58.8$$

Degree of Freedom. Between samples $v_1 = (k-1) = 1-1 = 1$, Within samples $v_2 = N-k = 45-2 = 43$

Analysis of Variance Table.

Source	S.S.	d.f.	M.S.	F	$F_0.5(1.43) = 4.06$
Between Samples	16	1	16	$ds\frac{16}{1.3674}$	$F_{.01}(1,43) = 7.23$
Within Samples	58.8	43	1.3674	=11.70	
Total	74.8	44			

Since $F_{cal} > F_{.01}(1,43)$, hence the difference is significant.

13.4 TWO WAY CLASSIFICATION

When the classification is based on two criteria, to study the effects of two kinds of treatments, then it is called two way classification. for example, To test simultaneous by the effect of fertilizer and kinds of seeds on agricultural produce, to test the effect of machines and laborers on industrial produce. The analysis of two way classification data is called two way analysis of variance or Analysis of variance in Two way classification.

13.5 TECHNIQUES OF TWO WAY ANALYSIS OF VARIANCE

1. Let us call the two criteria as row and column. then total sum of squares is partitioned into three components:

$$\begin{aligned} \text{Total sum of squares } &= \text{sum of squares between column} \\ &+ \text{sum of squares between row} \\ &+ \text{Residual (or Error) sum of squares} \\ \text{i.e.,} \qquad TSS &= SSC + SSR + SSE \end{aligned}$$

2 **Calculation :**

(i) The sum of all the observations, column wise, row wise and grand total. $\sum x_c, \Sigma x_v, \Sigma x$.

Let there be r rows and c columns. So that there are $r+c$ cells having one observation each. The number of observations in each row = n_c or c. The number of observations in each column = n_r or r.

Total number of observations = $r \times c = N$, say

(ii) **Correction factor :** $C.F. = \dfrac{T^2}{N}$

(iii) Total sum of squares

$$TSS = \Sigma(x^2) - \frac{T^2}{N}$$

$$\begin{aligned} \text{where } \Sigma x^2 &= \text{Sum of the squares of all the observations} \\ &= \sum(x_1^2 + x_2^2 + \ldots + x_n^2) \end{aligned}$$

(iv) **Sum of squares between columns:**

$$SSC = \Sigma\frac{(\Sigma x_c)^2}{r} - \frac{T^2}{N}$$

where $\Sigma\frac{(\Sigma x_c)^2}{r}$ = Sum of the squares of the totals of all the columns divided by the number of observation in each column

$$= \Sigma\frac{(\Sigma x_{c_1})^2}{r} + \Sigma\frac{(\Sigma x_{c_2})^2}{r} + \ldots + \Sigma\frac{(\Sigma x_{c_k})^2}{r}$$

(v) **Sum of squares between rows :**

$$SSR = \Sigma\frac{(\Sigma x_r)^2}{c} - \frac{T^2}{N}$$

where $\Sigma\frac{(\Sigma x_r)^2}{c}$ = Sum of the squares of the totals of all the rows divided by the number of observations in each row

$$= \frac{(\Sigma x_{r_1})^2}{c} + \frac{(\Sigma x_{r_2})^2}{c} + \ldots.. + \frac{(\Sigma x_{r_r})^2}{c}$$

(vi) **Residual sum of squares:** $SSE = TSS - (SSC + SSR)$
(vii) **Number of degrees of freedom :**
Degrees of freedom for total sum of squares = $r_c - 1 or N - 1$
Degrees of freedom for between column sum of squares = $c - 1$
Degree of freedom for between row sum of squares = $r - 1$
Degree of freedom for residual = $(c-1)(r-1)$
(ix) **Mean sum of squares :**

$$MSSC = \frac{SSC}{c-1},\ MSSR = \frac{SSR}{r-1},\ MSSE = \frac{SSE}{(c-1)(r-1)}$$

(x) **F-Ratio :**

$$F \text{ for columns } = \frac{MSDC}{MSSE},\quad F \text{ for rows } = \frac{MSSR}{MSSE}$$

Analysis of Variance Table.

Source	S.S.	d.f.	M.S.S	F
Columns	$\frac{\Sigma(\Sigma x_c)^2}{n_c} - \frac{T^2}{N} = S_1^2$	$c-1$	$S_1^2 \div (c-1)$	$\frac{MSSC}{MSSE}$
Rows	$\frac{\Sigma(\Sigma x_r)^2}{n_r} - \frac{T^2}{N} = S_2^2$	$r-1$	$S_2^2 \div (c-1)$	$\frac{MSSR}{MSSE}$
Error	$SST - S_1^2 + S_2^2 = S_3^2$	(c-1)(r-1)	$\frac{S_3^2}{(c-1)(r-1)}$	
Total	$\Sigma(x_1^2 + x_2^2 + \ldots x_k^2) - \frac{T^2}{N}$	$(c-1)$		

■ EXAMPLE 13.5

Four observers determine the moisture content of samples of a powder, each man taking a sample from each of six consignments. Their assessments are given below. Analyze this data and discuss whether there is any significant difference between consignment or between observers?

Observer	Consignment					
	1	2	3	4	5	6
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	12	13	11	14	12	10

Solution: Subtract 10 from each observation:

Observer	Consignment Column						
	1	2	3	4	5	6	Total
1	-1	0	-1	0	1	1	0
2	2	1	-1	1	0	0	3
3	1	0	0	2	1	0	4
4	2	3	1	4	2	0	12
Total	4	4	-1	7	4	1	19

Correction Factor

$$C.F. = \frac{T^2}{N} = \frac{19^2}{24} = \frac{361}{24} = 15.04$$

Total sum of squares

$$\begin{aligned} TSS &= \left[(-1)^2+0^2+(-1)^2+0^2+1^2+1^2\right]+\left[2^2+1^2+(-1)^2+0^2+1^2+1^2\right] \\ &\quad +\left[1^2+0^2+0^2+2^2+1^2+0^2\right]+\left[2^2+3^2+1^2+4^2+2^2+0\right]-C.F. = 35.96 \end{aligned}$$

Sum of squares between consignments:

$$\begin{aligned} SSC &= \frac{1}{4}(\Sigma c_1)^2+(\Sigma c_2)^2+(\Sigma c_3)^2+(\Sigma c_4)^2+(\Sigma c_5)^2+(\Sigma c_6)^2-C.F. \\ &= \left[\frac{4^2+4^2+(-1)^2+7^2+4^2+1^2}{4}\right]-15.04 = 9.71 \end{aligned}$$

Sum of squares between observers :

$$\begin{aligned} SSR &= \frac{(\Sigma r_1)^2+(\Sigma r_2)^2+(\Sigma r_3)^2+(\Sigma r_4)^2}{6}-C.F. \\ &= \frac{0+9+16+144}{6}-15.04 = 13.13 \end{aligned}$$

Error sum of squares

$$\begin{aligned} SSE &= TSS-(SSC+SSR) \\ &= 35.96-(9.71+13.13 = 35.96-22.84 = 13.12 \end{aligned}$$

Analysis of Variance Table.

Source	S.S.	d.f.	MSS	F
Consignment	9.71	6-1=5	$\frac{9.71}{5}=1.940$	$\frac{1.94}{0.87}=2.23<F_{.05}$
Observer	13.13	4-1=3	$\frac{13.13}{3}=4.38$	$\frac{4.38}{0.87}=5.03>F_{.05}$
Error	13.12	23-8=15	$\frac{13.12}{15}=0.87$	-
Total	35.96	25-1=23	-	-

Conclusion: $F_{.05}(v_1=5, v_2=15)=2.901$, $F.05v_1=3, v_2-15=3.287$. The difference among consignments is not significant. The difference among observers is significant.

Problems

13.1 In an experiment conducted on a farm in a certain village of Rajasthan, the following information was collected regarding the yield in quintal per acre of 6 plots of wheat. Three out of these six plots produced Sharbati wheat and the rest three produced Australian wheat. Set up an analysis of variance table and find out if the variety differences are significant. Use values of F at 5 level of significance.

Variety	Plot yield in quintal		
Sharbati	10	15	11
Australian	13	12	17

$F_{0.05}(v_1=4, v_2=1)=7.71$

13.2 A special fertilizer was used on four fields A,B,C and D. Four beds were made in each field. The yield statistics of all the four fields (A,B,C and D) are given below. Find whether the difference between the averages of yields of fields is significant or not. [Value of F at 5 level of significance for $v_2=3$ and $v_1=12$ is 8.74]

Yields of Fields			
A	B	C	D
8	9	3	3
12	4	8	7
1	7	2	8
3	1	5	2

13.3 The table below gives the yields in quintal of four plots each of three varieties of wheat. Is there a significant difference between the mean yield of the three varieties ? [$F_{.05}$ for degrees of freedom $v_1=2$ and $v_2=9$ is 4.26

Plot No.	Varieties		
	A	B	C
1	10	9	4
2	6	7	8
3	7	7	6
4	9	5	6

13.4 The following data relating to the output of sugarcane in three varieties obtained from 4 plots of each variety. Setup a table of analysis of variance and also find out whether the [$F_{.05}$ for degrees of freedom, $v_1 = 9$ and $v_2 = 2$ is 19.38

Varieties	Plots			
	1	2	3	4
X	20	16	12	12
Y	14	14	16	20
Z	12	12	14	18

13.5 Following figures relate to the production of rice in kilogram of 3 varieties shown in 9 plots :

Type A	Type B	Type C
14	14	18
16	13	16
	15	19
		19

Is there any significant difference in the 3 types of plots.

13.6 To test the efficacy of a particular manure four fields A,B,C and D, were selected. Each field was cut in 4 plots and the manure was used in them. The yields obtained from these plots are given below. Find out whether there is or not any significant difference between the mean yields of the fields (the value of F for 3 and 9 degrees of freedom for error at 5 level of significance is 3.49)

Plots	Yields from the Fields			
	A	B	C	D
1	12	8	11	10
2	14	12	13	12
3	13	11	11	8
4	17	13	13	10

13.7 The following figures relate to production (in kilogram) of three varieties A,B and C of wheat shown in 12 plots. Is there any significant difference in the production of these varieties?

A	14	16	18		
B	14	13	15	22	
C	18	16	19	19	20

13.8 The following data relate to the output of three varieties of rice obtained from four plots. Set up a table of analysis of variance and also find out whether the difference between the yields of the three varieties is significant or not?

Varieties	Plots			
	1	2	3	4
X	12	16	12	20
Y	12	14	18	12
Z	16	14	20	14

13.9 The following table gives the average monthly sale (in thousand of rupees) of four salesmen in the different types of territories:

Territory	Salesman A	B	C	D	Territory Total
X	6	4	8	6	24
Y	7	6	6	9	28
Z	8	5	10	9	32
Salesman Total	21	15	24	24	84

Set up an analysis of variance table for the data given above.

13.10 Below are given the yields of three strains of wheat planted in five blocks of three plots each under a completely randomized design. All the fifteen plots are of equal area. Perform the analysis of variance to test whether the strains are significantly different with regard to yield. Ignore variations between blocks.

Strain	Blocks I	II	III	IV	V
A	20	21	23	16	20
B	18	20	17	15	25
C	25	28	22	28	32

13.11 Five doctors each tests five treatments for a certain disease and observe the number of days each patient takes to recover. The following table gives the recovery times in days corresponding to each doctor and each treatment.

Doctors	Treatments Method 1	2	3	4	5
A	11	15	24	20	21
B	12	16	25	18	22
C	10	13	21	17	20
D	9	14	18	18	21
E	13	16	20	16	23

Carry out the analysis of variance and test whether there is any significant difference (a) between the doctors, and (b) between the treatments.

CHAPTER 14

TIME SERIES AND FORECASTING

Time series analysis is the basis for understanding past behaviour, evaluating current accomplishment, planning future operations. It is also used for comparing the components of different time series. It is used to determine the patterns in the data of the past over a period of time and extrapolate the data into the future. Previous performances are studied to forecast future activity.

Prof. W.Z. Hirsch writes,

> A main objective in analyzing a time series is to understand, interpret and evaluate changes in economic phenomena in the hope of more correctly anticipating the course of the events.

In short, following are the use of time series analysis:

1. Analysis of past behaviour: Time series analysis helps in understanding the past behaviour of the factors which are responsible for the variations.
2. Estimates for the future : The understanding of the past behaviour and projecting the past trends are extremely helpful in predicting the future behaviour.
3. Forecasting: Time series study helps in forecasting and planning future operations.
4. Evaluation of performance: Time series analysis helps in evaluating current accomplishments.
5. Comparison: Time series analysis help in comparison among various time series. Interpreting of the variations that how they are related with each other, net effect of their interaction and also with the similar changes in other time series data.
6. Estimation of trade cycles: Time series analysis helps in the estimation of trade cycles on the basis cyclical fluctuations which helps the businessman to plans and regulate their activities.

14.1 COMPONENTS OF TIME SERIES

The forces affecting time series data generate certain movements or fluctuations in a time series. Such characteristics movements or fluctuations of time series are called components of a time series. The components of a time series may be classified into different categories on the basis of the operations forces. There are four components of a time series.

1. Secular Trend or Trend (T)
2. Seasonal Variations (S)
3. Cyclic Variations (C)
4. Irregular (or Random) Variations (I)

The components of a time series may or may not occur at the same time.

14.1.1 Trend or Secular Trend

The component of a time series which is responsible for its general behaviour (i.e. general long term movement) over a fairly long period of time as a result of some identifiable influences is called the secular trend or trend. It is a smooth, regular and long term tendency of a particularly activity to grow or decline. The trend may be upward as well as downward. If the series neither increases nor decreases over a long term (usually a minimum of 15 to 20 years) then series is said without trend (or nor-trend) or with a constant trend:

It is not necessary that the trend should be in the same direction through out the given period. The time series may be increasing slowly or increasing fast or may be decreasing at various rates or may remain relatively constant. Some time series reverse their trend from growth to decline or from decline to growth over a period of time. In brief, the movements which exhibit persistent growth or decline in long period of time are known as secular trend. The term 'long period of time' is a relative concept which is influenced by the characteristic of the series.

The formation of rocks is a particular example of secular trend. Declining death rate is an example of downward trend; population growth is an example of upward trend.

Mathematically trend may be linear or nonlinear.

14.1.2 Seasonal Variations

The movements that are regular and periodic in nature not exceeding a year are called seasonal variations.

The movements that are regular and periodic in nature not exceeding a year are called seasonal variations.

The forces that are responsible for seasonal variations are:- (i) Natural factors (or climate) (ii) Man-made conventions (Holidays, Festivals, etc.)

For example, crops are sown and harvested at certain times every year and demand for labor goes up during sowing and harvesting seasons, demands of woolen cloths goes up in winter, prices increase during festivals, with draws from banks are heavy on first week of a month, the number of letters posted on Saturday is larger, etc.

14.1.3 Cyclic Variations

The oscillatory movements (or swings or patterns) whose period of oscillation is more than one year are called *cyclic variations* one complete period is called a *cycle*. In economic and business series they correspond to the business cycle and take place as a result of economic booms or depressions.

It is a matter of common knowledge that almost all economic and business activities have four distinct phases: Prosperity, Decline, Depressions, Recovery

14.1.4 Irregular (or Random) Variations

Irregular (or Random) Variations do not exhibit any definite pattern and there is no regular period of their occurrence. these are accidental changes which are purely random and unpredictable. For example, variations due to earthquake, war etc. Normally these variations are short term but sometimes their effects are severe.

14.2 MATHEMATICAL MODELS OF THE TIME SERIES

For an analysis of time series, the traditional or classical method is to suppose some type of relationship among the components of a time series. The two relationships often called *Models of Time Series* are as follows:

1. Additive Model: Here, we assume that the various components of a time series are additive and the values are the sum of the four components. Symbolically,

$$Y = T + S + C + I$$

   where Y= Observed value in the time series, T= Trend component, S= Seasonal variations component, C= Cyclical variations component, I= Irregular variations component

2. Multiplicative Model: Here we assume that any particular value in a time series is the product of the four components. Symbolically,

$$Y = T \times S \times C \times I$$

   where Y= Observed value in the time series, T= Trend component, S= Seasonal variations component, C= Cyclical variations component, I= Irregular variations component

14.3 MEASUREMENT OF SECULAR TREND

Trends are measured by the following methods:

1. Free hand graphic method
2. Method of semi averages
3. Method of moving averages
4. Method of least squares

Here we are interested in only method of semi averages and method of moving averages.

14.3.1 Method of Semi Averages

The methods of fitting a linear trend with the help of semi-average method are as follows:

1. *The number of years is even*: The data of the time series are divided into two equal parts. The total of the items in each as the part is done and it is then divided by the number of items to obtain arithmetic means of the each part. Each average is then plotted at mid-point of the interval from which it has been computed. A straight line is drawn passing through these points. This is the required trend line.
2. *The number of years is odd*: When the number of years is odd, the value of the middle year is omitted to divide the time series into equal parts. Then follow the same procedure as for even numbers.

Mathematical form of the trend line. The straight line passing through the two points $(t_1 - y_1)$ and $(t_2 - y_2)$,

$$y - \overline{y_1} = \frac{\overline{y_2} - \overline{y_1}}{t_2 - t_1}(t - t_1)$$

The value $\dfrac{\overline{y_2} - \overline{y_1}}{t_2 - t_1}$ is the annual increase in y for a unit change in t.

■ EXAMPLE 14.1

Determine the trend for the following data by semi average method:

Year	1991	1992	1993	1994	1995	1996
Sales(y)	60	80	100	80	120	100

Solution: Here, number of data is even. Divide data in two part and take mean($\bar{y}_1$,$\bar{y}_2$) of each part as

$$\bar{y}_1 = \frac{60 + 80 + 100}{3} = \frac{240}{3} = 80$$

Mean year for first part, $t_1 = 1992$

$$\bar{y}_2 = \frac{80 + 120 + 100}{3} = \frac{300}{3} = 100$$

Mean year for first part, $t_2 = 1995$

Trend line

$$y - \overline{y_1} = \frac{\overline{y_2} - \overline{y_1}}{t_2 - t_1}(t - t_1)$$

$$y - 100 = \frac{100 - 80}{1995 - 1992}(t - 1995)$$

$$\Rightarrow y = \frac{20}{3}(t - 1980)$$

On graph, we can plot line through points (1992,80)(1995,100). □

■ EXAMPLE 14.2

Determine the trend for the following data by semi average method:

Year	1991	1992	1993	1994	1995	1996	1997
Sales(y)	60	80	100	80	120	100	80

Solution: Here, number of data is odd. Therefore leave middle data, i.e., 80 corresponding to 1994. Now divide data in two part and take mean($\bar{y}_1$,$\bar{y}_2$) of each part as

$$\bar{y}_1 = \frac{60 + 80 + 100}{3} = \frac{240}{3} = 80$$

Mean year for first part, $t_1 = 1992$

$$\bar{y}_2 = \frac{120 + 100 + 80}{3} = \frac{300}{3} = 100$$

Mean year for first part, $t_2 = 1996$

Trend line

$$y - \overline{y_1} = \frac{\overline{y_2} - \overline{y_1}}{t_2 - t_1}(t - t_1)$$

$$y - 100 = \frac{100 - 80}{1996 - 1992}(t - 1996)$$

$$\Rightarrow \; y = 5(t - 1976)$$

On graph, we can plot line through points (1992,80)(1996,100). Plot yourself. □

14.3.2 Method of Moving Average

To determine trend with the following steps is called the method of moving averages:

1. We take the first m consecutive values of the series and calculate these average, m is a number to be decided by us and is called the time-interval or period for computing the moving average. This average is taken as the trend value for the time failing at the middle of the period covered in the computation of average. When the time id odd it is all right. But when the time period is even, the moving average fall mid way between the two observations, because the mid term will not coincide with a point of the series. To avoid this difficulty we calculate further two-item moving averages so that the mid-points now may coincide with the given time periods. This process is known as centering of the moving averages.
2. We drop the first observation in the series and include the $(m+1)$th observation and calculate their average and place it at the middle of the period covering second to $(m+1)$th values.
3. Now leave the first two values and include the $(m+1)$th value and calculate their average and place it a the middle of the period covering third to $(m+2)$th values. This process will go on till the and of the series. Symbolically, suppose the time series is given
4. Time:

Time	t_1	t_2	...	t_m	...	t_n
Value	y_1	y_2	...	y_m	...	y_n

Let $m = 3$ (Odd case), then

$$\text{First moving average } v_1 = \frac{y_1 + y_2 + y_3}{3} \text{ against } t_2$$

$$\text{Second moving average } v_2 = \frac{y_2 + y_3 + y_4}{3} \text{ against } t_2$$

$$\text{Third moving average } v_3 = \frac{y_3 + y_4 + y_5}{3} \text{ against } t_2$$

The quantities $v_1, v_2, v_3, ..$ etc are called 3-yearly moving average.
Now again let $m = 4$ (Even case)

$$\text{First moving average } v_1 = \frac{y_1 + y_2 + y_3 + y_4}{4} \text{ against } t_2$$

$$\text{Second moving average } v_2 = \frac{y_2 + y_3 + y_4 + y_5}{4} \text{ against } t_2$$

$$\text{Third moving average } v_3 = \frac{y_3 + y_4 + y_5 + y_6}{4} \text{ against } t_2$$

The quantities $v_1, v_2, v_3, ..$ etc are called 4-yearly moving average. In even case we centered moving averages, therefore centered moving averages are $\frac{v_1 + v_2}{2}$, $\frac{v_2 + v_3}{2}$, etc.

The technique of moving averages is used to eliminate the fluctuations and give only the general trend of the series.

EXAMPLE 14.3

Use the following data to compute a 3-year moving average for all available years. Also determine the trend.

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Production	21	22	23	25	24	22	25	26	27	26

The first average is computed for the first 3-year as follows:

$$\text{Moving average(for year 1995,1996,1997)} = \frac{21+22+23}{3} = 22.00$$

Similarly, the moving average for the next 3-year is

$$\text{Moving average(for year 1996,1997,1998)} = \frac{22+23+25}{3} = 23.33$$

Other moving averages are as follows:

Year	Production	3-Year Moving Total	3-Year Moving Average (Trend Values)
1995	21		
1996	22	21+22+23=66	66/3=22.00
1997	23	22+23+25=70	70/3=23.33
1998	25	72	24.00
1999	24	71	23.67
2000	22	71	23.67
2001	25	73	24.33
2002	26	78	26.00
2003	27	26+27+26=79	26.33
2004	26		

EXAMPLE 14.4

Assume a four year cycle and calculate the trend from the following data:

Year	1995	1996	1997	1998	1999	2000	2001	2002
Production	464	515	518	467	502	540	557	571

The first average is computed for the first 4-year as follows:

$$\text{Moving average(for year 1995,1996,1997,1998)} = \frac{464+515+518+467}{4} = 491.00$$

Similarly, the moving average for the next 4-year is

$$\text{Moving average(for year 1996,1997,1998,1999)} = \frac{515+518+467+502}{4} = 500.50$$

Other moving averages are as follows:

Year Year	Production Production	4-Year Moving Total	4-Year Moving Average	4-Year Moving Average Centered (Trend Values)
1995	464			
1996	515			
		464 + 515 +518 + 467=1964	1964/4=491.00	
1997	518			=(491.00+500.50)/2=495.75
		2002	500.50	
1998	467			503.62
		2027	506.75	
1999	502			511.62
		2066	516.50	
2000	540			529.50
		2170	542.50	
2001	557			
2002	571			

14.4 MEASUREMENT OF SEASONAL VARIATION

In order to isolate and identify seasonal variations, we first eliminate as far as possible the effect of trend, cyclical variations and irregular fluctuations on the time series. The main methods of measuring seasonal variations are:

1. Simple Average Method

2. Ratio to Moving Average Method

3. Link Relative Method

4. Ratio to Trend Method

We will discuss only two methods here: Simple Average Method and Ratio to Moving Average Method.

14.4.1 Simple Average Method (Monthly And Quarterly) Seasonal Method

1. The time series is arranged by years and months (or quarters).

2. The average for each month (or quarter) is obtained (the average may be mean or median). In general, we take mean if not specified otherwise.

3. Now compute seasonal index for each month (or quarter) as

$$\frac{\text{Average of } m\text{th Month(or quarter)}}{\text{Average of monthly(or quarterly) averages}} \times 100$$

■ **EXAMPLE 14.5**

Find the seasonal indices from the data regarding monthly consumption of wheat in a town

Months	1971	1972	1973	1974	1975
Jan.	44	47	48	50	51
Feb.	45	48	44	58	50
March	39	43	43	45	50
April	42	44	47	52	60
May	41	45	48	51	55
June	40	46	47	52	55
July	48	53	55	57	62
Aug.	49	44	56	55	51
Sept	42	45	47	49	57
Oct.	45	48	51	52	64
Nov.	41	44	48	52	60
Dec.	50	55	57	63	60

Solution:

Month	1971	1972	1973	1974	1975	Average of 5 years	Seasonal-index
Jan	44	47	48	50	51	48	96
Feb	45	48	44	58	50	49	98
Mar	39	43	45	45	50	44	88
April	42	44	45	52	60	49	98
May	41	45	48	51	55	48	96
June	40	46	47	52	55	48	96
July	39	53	55	57	62	55	110
Aug.	49	44	56	55	51	51	102
Sept.	42	45	47	49	57	48	96
Oct.	45	48	51	52	64	52	104
Nov.	41	40	48	52	60	59	98
Dec.	50	55	57	63	70	59	118
Total						600	
Average of Averages						50	

$$\text{Seasonal index}=\frac{\text{Average of mth Month}}{\text{Average of monthly averages}}\times 100$$

$$\text{Seasonal index for Jan.} = \frac{\text{Average of Jan. Month}}{\text{Average of monthly averages}}\times 100$$

$$= \frac{48}{50}\times 100 = 96$$

Similarly the seasonal indices of other months are obtained. The required seasonal indices are shown in the last column of the table. □

■ EXAMPLE 14.6

Compute seasonal indexes by the average percentage method and obtain deseasonalized values for the following data:

Quarter	Years				
	2004	2005	2006	2007	2008
I	45	48	49	52	60
II	54	56	63	65	70
III	72	63	70	75	84
IV	60	56	65	72	66

Calculation for quarterly averages are given in table:

$$\text{Average of quarterly averages} = \frac{50.8+61.6+72.8+63.8}{4}$$

Quarter	Years					Total	Average
	2004	2005	2006	2007	2008		
I	45	48	49	52	60	254	50.8
II	54	56	63	65	70	308	61.6
III	72	63	70	75	84	364	72.8
IV	60	56	65	72	66	319	63.8
			Grand Average→				62.25

Therefore,

$$\text{Seasonal Index for quarter I} = \frac{50.8}{62.25} \times 100 = 81.60$$
$$\text{Seasonal Index for quarter II} = \frac{61.6}{62.25} \times 100 = 98.95$$
$$\text{Seasonal Index for quarter III} = \frac{72.8}{62.25} \times 100 = 116.94$$
$$\text{Seasonal Index for quarter IV} = \frac{63.8}{62.25} \times 100 = 102.48$$

Deseasonalized Values.

$$\text{Deseasonalized Values} = \frac{\text{Actual quarterly value}}{\text{Seasonal index of corresponding quarter}} \times 100$$

Thus deseasonalized value for year 2004 and quarter first is

$$\text{Deseasonalized Values} = \frac{45}{81.60} \times 100 = 55.14$$

Other values are given in table

Quarter	Years				
	2004	2005	2006	2007	2008
I	55.14	58.82	60.00	63.72	73.52
II	54.57	56.59	63.66	65.68	70.74
III	61.57	53.87	59.85	64.13	71.83
IV	58.54	54.64	63.42	70.25	64.40

14.4.2 Moving Average Method

In this method moving averages of the data are computed. If the data are monthly then 12-monthly moving average, if they are quarterly, then 4-quarterly moving averages will be computed. In both the cases a time periods of moving averages are even, hence these moving averages are to be centered.

Under additive model, from each original value, the corresponding moving averages is deducted to find out short-time fluctuations. By preparing a separate table, monthly (or quarterly) short-time fluctuations are added for each month (or quarter) over all the years and their averages is obtained. These averages are known as seasonal variations for each month or quarter. If one wants to isolate/measure irregular variations, the mean of the respective month or quarter is deducted from the short-time fluctuations.

■ **EXAMPLE 14.7**

Find out seasonal fluctuations by the method of moving averages:

Years	1st quarter	2nd quarter	3rd quarter	4th quarter
1976	30	40	36	34
1977	34	52	50	44
1978	40	58	54	48
1979	54	76	58	62
1980	80	92	86	82

Solution:

Years	Quarter	Quarter data (y)	Quarter Total	Total centered	Quality Moving Average (T)	Short time fluctuations	Seasonal Variations(S)
(i)	(ii)	(iii)	(iv)	(v)	(vi)=[v/8]	(vii)=[(iii)-(vi)]	(viii)
1976	I	30	-	-	-	-	-
	II	40	140	-	-	-	-
	III	36	144	284	35.5	+0.5	-0.2
	IV	34	156	300	37.5	-3.5	-3.9
1977	I	34	170	326	40.8	-6.8	-4.1
	II	52	180	350	43.8	+8.2	10.4
	III	50	186	366	45.8	+4.2	-0.2
	IV	44	192	378	45.3	-3.3	-5.9
1978	I	40	196	388	48.5	-8.5	-4.1
	II	58	200	396	49.5	+8.5	10.4
	III	54	214	414	51.8	+2.2	-0.2
	IV	48	232	446	55.8	-7.8	-5.9
1979	I	54	236	468	58.5	-4.5	-4.1
	II	76	250	486	60.8	+15.2	10.4
	III	58	276	526	65.8	-7.8	-0.2
	IV	62	292	568	71.0	-9.0	-5.9
1980	I	80	320	612	76.5	+3.5	-4.1
	II	92	340	660	82.5	+9.5	10.4
	III	86	-	-	-	-	-
	IV	82	-	-	-	-	-

Calculation of Average Seasonal Variations.

Years	Quarters I	II	III	IV
1976	–	–	+0.5	-3.5
1977	-6.8	+8.2	+4.2	-3.3
1978	-8.5	+8.5	+2.2	-7.8
1979	-4.5	+15.2	-7.8	-9.0
1980	+3.5	+9.5	–	—
Total	-16.3	41.4	-0.9	-23.6
Average	-4.1	10.4	-0.2	-5.9

□

Problems

14.1 Determine the trend values by three yearly moving averages:

Year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Variable	8	12	10	13	15	12	16	17	14	17

14.2 Calculate three yearly moving averages from the following data:

Year	1990	1991	1992	1993	1994
No.of Wage Earners	2,412	2,440	2,486	2,424	2,458
Year	1995	1996	1997	1998	
No.of Wage Earners	2,513	2,409	2,488	2,528	

14.3 Find 5-yearly moving averages for the following time series and represent them on the same graph along with the original data:

Year	1	2	3	4	5	6	7	8	9	10
Value	110	104	78	105	109	120	115	110	114	122

14.4 Plot the following data on a graph paper. calculate five yearly moving averages and show the trend on the graph paper:

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Index No	105	115	100	90	80	95	85	75	60	67

14.5 Index numbers from 1979 to 1993 are given in the following table. calculate trend value using 5 yearly moving averages:

Year	1979	1980	1981	1982	1983	1984	1985	1986
Index	100	138	187	222	245	139	134	229
Year	1987	1988	1989	1990	1991	1992	1993	
Index	192	260	182	247	311	191	261	

14.6 Assuming trend to be absent, calculate seasonal variations for the time-series data given below:

	Quarters			
Year	I	II	III	IV
1991	31	40	44	37
1992	40	32	42	40
1993	36	40	39	37
1994	35	30	40	33

CHAPTER 15

STATISTICAL QUALITY CONTROL

Statistica1 quality control (SQC) is the term used to describe the set of statistical tools used by quality professionals. Statistical quality control can be divided into two broad categories:

1. **Statistical process control** (SPC) involves inspecting a random sample of the output from a process and deciding whether the process is producing products with characteristics that fall within a predetermined range. SPC answers the question of whether the process is functioning properly or not.
2. **Acceptance sampling** is the process of randomly inspecting a sample of goods and deciding whether to accept the entire lot based on the results. Acceptance sampling determines whether a batch of goods should be accepted or rejected.

15.1 STATISTICAL PROCESS CONTROL METHODS

Statistical process control methods extend the use of descriptive statistics to monitor the quality of the product and process. As we have learned so far, there are common and assignable causes of variation in the production of every product. Using statistical process control we want to determine the amount of variation that is common or normal. Then we monitor the production process to make sure production stays within this normal range. That is, we want to make sure the process is in a state of control. The most commonly used tool for monitoring the production process is a control chart.

15.1.1 Control Charts

A control chart (also called process chart or quality control chart) is a graph that shows whether a sample of data falls within the common or normal range of variation. A control chart has upper and lower control limits that separate common from assignable causes of variation. The common range of variation is defined by the use of control chart limits.We say that a process is out of control when a plot of data reveals that one or more samples fall outside the control limits.

The *x* axis represents samples taken from the process over time. The *y* axis represents the quality characteristic that is being monitored. The *center line* (CL) of the control chart is the mean, or average, of the quality characteristic that is being measured. The *upper control limit* (UCL) is the maximum acceptable variation from the mean for a process that is in a state of control. Similarly, the *lower control limit* (LCL) is the minimum acceptable variation from the mean for a process that is in a state of control.

The upper and lower control limits on a control chart are usually set at $\pm 3\times$ (standard deviations) from the mean. If we assume that the data exhibit a normal distribution, these control limits will capture 99.74 percent of the normal variation. If we assume that the data exhibit a normal distribution, these control limits will capture 99.74 percent of the normal variation. Control limits can be set at $\pm 2\times$ (standard deviations) from the mean. In that case, control limits would capture 95.44 percent of the values. These control limits are are called as inner control limits and outer control limits for 95.45 percent level. These limits also called as warning limits.

15.2 TYPES OF CONTROL CHARTS

Control charts are one of the most commonly used tools in statistical process control. They can be used to measure any characteristic of a product, such as the weight of a cereal box, the number of chocolates in a box, or the volume of bottled water. The different characteristics that can be measured by control charts can be divided into two groups: variables and attributes.

A *control chart for variables* is used to monitor characteristics that can be measured and have a continuum of values, such as height, weight, or volume. A soft drink bottling operation is an example of a variable measure, since the amount of liquid in the bottles is measured and can take on a number of different values. Other examples are the weight of a bag of sugar, the temperature of a baking oven, or the diameter of plastic tubing.

A *control chart for attributes*, on the other hand, is used to monitor characteristics that have discrete values and can be counted. Often they can be evaluated with a simple yes or no decision. Examples include color, taste, or smell. The monitoring of attributes usually takes less time than that of variables because a variable needs to be measured (e.g., the bottle of soft drink contains 15.9 ounces of liquid). An attribute requires only a single decision, such as yes or no, good or bad, acceptable or unacceptable (e.g., the apple is good or rotten, the meat is good or stale, the shoes have a defect or do not have a defect, the light bulb works or it does not work) or counting the number of defects (e.g., the number of broken cookies in the box, the number of dents in the car, the number of barnacles on the bottom of a boat).

15.2.1 Control charts for variables

Control charts for variables monitor characteristics that can be measured and have a continuous scale, such as height, weight, volume, or width. When an item is inspected, the variable being monitored is measured and recorded. For example, if we were producing candles, height might be an important variable.We could take samples of candles and measure their heights. Two of the most commonly used control charts for variables monitor both the central tendency of the data (the mean) and the variability of the data (either the standard deviation or the range). Note that each chart monitors a different type of information. When observed values go outside the control limits, the process is assumed not to be in control. Production is stopped, and employees attempt to identify the cause of the problem and correct it

15.2.1.1 Mean (x-Bar) Charts A mean control chart is often referred to as an x-bar chart. It is used to monitor changes in the mean of a process. To construct a mean chart we first need to construct the center line of the chart. To do this we take multiple samples and compute their means. Usually these samples are small, with about four or five observations. Each sample has its own mean, $\overline{x_i}$. The center line of the chart is then computed as the mean of all N sample means, where N is the number of samples:

$$\overline{\overline{x}} = \frac{\sum x_i}{N}$$

To construct the upper and lower control limits of the chart, we use the following formulas:

$$\text{Upper control limit (UCL)} = \overline{\overline{x}} + 3\sigma_{\overline{x}}$$

$$\text{Lower control limit (LCL)} = \overline{\overline{x}} + 3\sigma_{\overline{x}}$$

where $\overline{\overline{x}}$ = the average of the sample means
$\sigma_{\overline{x}}$ = standard deviation of the distribution of sample means, computed as $\frac{\sigma_p}{\sqrt{n}}$
σ_p = population (process) standard deviation
n = sample size (number of observations per sample)

Another way to construct the control limits is to use the sample range as an estimate of the variability of the process. Remember that the range is simply the difference between the largest and smallest values in the sample. The spread of the range can tell us about the variability of the data. In this case control limits would be constructed as follows:

$$\text{Upper control limit (UCL)} = \overline{\overline{x}} + A_2\overline{R}$$

$$\text{Lower control limit (LCL)} = \overline{\overline{x}} + A_2\overline{R}$$

where $\overline{\overline{x}}$ average of the sample means
$\overline{R}$ average range of the samples
A_2 quality control factor.

■ EXAMPLE 15.1

A quality control inspector at the Sprite soft drink company has taken ten samples with four observations each of the volume of bottles filled. The data and the computed means are shown in the table. If the standard deviation of the bottling operation is 0.13 ounces, use this information to develop control limits of three standard deviations for the bottling operation.

Sample	Observations(bottle volume in ml.)				Average	Range
Numbers	I	II	III	IV	$\overline{x}$	R
1	15.85	16.02	15.83	15.93	15.91	0.19
2	16.12	16.00	15.85	16.01	15.99	0.27
3	16.00	15.91	15.94	15.83	15.92	0.17
4	16.20	15.85	15.74	15.93	15.93	0.46
5	15.74	15.86	16.21	16.10	15.98	0.47
6	15.94	16.01	16.14	16.03	16.03	0.20
7	15.75	16.21	16.01	15.86	15.96	0.46
8	15.82	15.94	16.02	15.94	15.93	0.20
9	16.04	15.98	15.83	15.98	15.96	0.21
10	15.64	15.86	15.94	15.89	15.83	0.30
Total					159.44	2.93

Solution: The center line of the control data is the average of the samples:

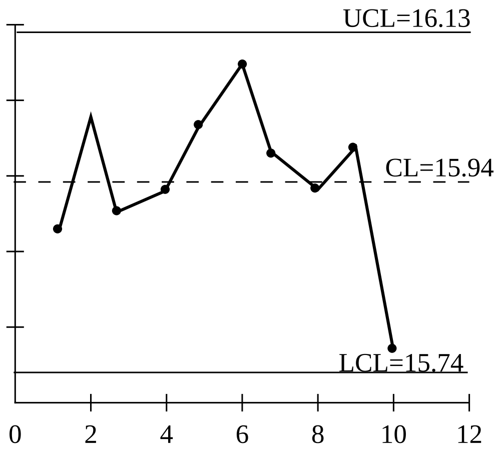
$$\overline{\overline{x}} = \frac{\sum x_i}{N} = \frac{159.44}{10} = 15.94$$

The control limits are

$$\text{Upper control limit (UCL)} = \bar{\bar{x}} + 3\sigma_{\bar{x}} = 15.94 + 3 \times \frac{0.14}{\sqrt{4}} = 16.13$$

$$\text{Lower control limit (LCL)} = \bar{\bar{x}} + 3\sigma_{\bar{x}} = 15.94 - 3 \times \frac{0.14}{\sqrt{4}} = 15.74$$

The resulting control chart is: □



■ EXAMPLE 15.2

The following table gives the average daily production figures for 20 months each of 25 working days . Given that the population standard deviation of daily production is 35 units , draw a control chart for the mean:

210	205	210	212	211	209	219	204	212	209
212	215	208	214	210	204	211	211	203	211

Solution: Given: $n = 25, k = 20, \sigma_p = 35$

$$\text{Central Line (CL)} = \bar{\bar{X}} = \frac{\bar{X_1} + \bar{X_2} + \bar{X_3} + \dots\dots\dots + \bar{X_k}}{k}$$

$$= \frac{4200}{20} = 210$$

$$UCL = \bar{\bar{X}} + 3\frac{\sigma_p}{\sqrt{n}} = 210 + (3 \times \frac{35}{\sqrt{25}}) = 231$$

$$LCL = \bar{\bar{X}} - 3\frac{\sigma_p}{\sqrt{n}} = 210 - (3 \times \frac{35}{\sqrt{25}}) = 189$$

□

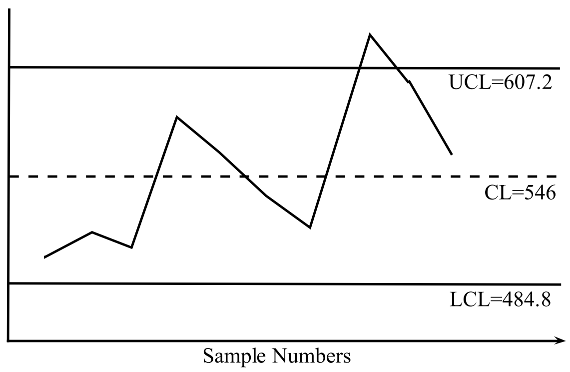
■ EXAMPLE 15.3

Mean and Range of 10 random samples of size 4 were:

Sample	1	2	3	4	5	6	7	8	9	10
Mean ($\bar{X}$)	494	508	500	582	552	538	514	614	600	558
Range (R)	95	125	100	91	68	65	148	28	37	80

Construct a Control Chart for Mean ($\bar{X}$ chart) and determine whether the process is within control or not ?(Given: Value of $d_2$for size of sample 4 is 2.059.)

Solution:



$$\bar{\bar{X}} = \frac{\sum \bar{X}}{k} = \frac{494+508+500+582+552+538+514+614+600+558}{10}$$
$$= \frac{5460}{10} = 546$$

$$\bar{R} = \frac{\sum R}{k} = \frac{95+128+100+91+68+65+148+28+37+80}{10}$$
$$= \frac{840}{10} = 84$$

$$\sigma_p = \frac{\bar{R}}{d_2} = \frac{84}{2.059} = 40.8$$

C.L. $= \bar{\bar{X}} = 546$
UCL $= \bar{\bar{X}} + 3\frac{\sigma_p}{\sqrt{n}} = 546 + (\frac{3*40.8}{\sqrt{4}}) = 546 + (\frac{3*40.8}{2}) = 546 + 61.2 = 607.2$
LCL $= \bar{\bar{X}} - 3\frac{\sigma_p}{\sqrt{n}} = 546 - 61.2 = 484.8$ □

15.2.1.2 Range (R) charts Range (R) charts are another type of control chart for variables. Whereas x-bar charts measure shift in the central tendency of the process, range charts monitor the dispersion or variability of the process. The method for developing and using R-charts is the same as that for x-bar charts. The center line of the control chart is the average range, and the upper and lower control limits are computed as follows: where values for D4 and D3 are obtained from Table 15.1..

$$CL = \bar{R}$$
$$UCL = D_4\bar{R}$$
$$LCL = D_3\bar{R}$$

Table 15.1. Factors for three-sigma control limits

Sample Size n	Factor for $\bar{x}$-Chart A_2	Factor for R-Chart D_3	D_4
2	1.88	0	3.27
3	1.02	0	2.57
4	0.73	0	2.28
5	0.58	0	2.11
6	0.48	0	2
7	0.42	0.08	1.92
8	0.37	0.14	1.86
9	0.34	0.18	1.82
10	0.31	0.22	1.78
11	0.29	0.26	1.74
12	0.27	0.28	1.72
13	0.25	0.31	1.69
14	0.24	0.33	1.67
15	0.22	0.35	1.65
16	0.21	0.36	1.64
17	0.2	0.38	1.62
18	0.19	0.39	1.61
19	0.19	0.4	1.6
20	0.18	0.41	1.59
21	0.17	0.43	1.58
22	0.17	0.43	1.57
23	0.16	0.44	1.56
24	0.16	0.45	1.55
25	0.15	0.46	1.54

■ EXAMPLE 15.4

Draw a control chart for the range R from the following data relating to 20 samples, each of size 5: Only the central line and the upper and lower control limits may be drawn in the chart. The sum of the 20 sample ranges in 410 inch. Given $d_3 = 0$ and $d_4 = 2.11$.

Solution: Given: $n = 5,\ k = 20,\ \sum R = 410$

$$CL = \bar{R} = \frac{\sum R}{k} = \frac{410}{20} = 20.5$$

$$UCL = d_4\bar{R} = 2.11 \times 20.5 = 43.26$$

$$LCL = D_3\bar{R} = 0$$

□

■ EXAMPLE 15.5

Range of 10 random samples of size 5 were: $22, 60, 13, 6, 21, 18, 20, 10, 40, 10$. Construct a control chart for Range (R-chart) and determine whether the process is within control or not. (Give value of D_3 for size of sample 5 is 0 and of $D_4 = 2.115$).

Solution:

$$CL = \bar{R} = \frac{\sum R}{k} = \frac{220}{10} = 22$$

$$UCL = D_4.\bar{R} = 2.115 \times 22 = 46.53$$

$$LCL = D_3.\bar{R} = 0 \times 22 = 0$$

Process is out of control at sample number 2 , because the point R_2 is outside the control limits. □

■ EXAMPLE 15.6

The observed values of 15 samples of size 4 are as follows:

32,20,30,14;42,36,52,50;25,16,52,63;22,33,34,23;28,30,27,31;
30,32,26,16;34,30,28,32;11,21,20,16;11,22,28,31;36,31,35,26;
34,16,37,25;28,36,51,53;16,35,32,37;35,36,37,24;17,28,14,13

Compute the control limits for $\bar{X}$ and R-charts.

Solution:

$$\bar{X} = \frac{\sum \bar{X}}{N} = \frac{445}{15} = 29.67$$

$$\bar{R} = \frac{\sum \bar{R}}{N} = \frac{253}{15} = 16.87$$

For $\bar{X}$-Chart:

$$UCL = \bar{\bar{X}} + A_2\bar{R} = 29.67 + 0.73 \times 16.87 = 41.99$$

$$CL = \bar{\bar{X}} = 29.67$$

$$LCL = \bar{\bar{X}} - A_2\bar{R} = 29.67 - 0.73 \times 16.87 = 17.35$$

Samples	Observations				Average ($\bar{x}$)	Range (R)
1	32	20	30	14	24	18
2	42	36	52	50	45	16
3	25	16	52	63	39	47
4	22	33	34	23	28	12
5	28	30	27	31	29	3
6	30	32	26	16	26	16
7	34	30	28	32	31	6
8	11	21	20	16	17	10
9	11	22	28	31	23	20
10	36	31	35	26	32	10
11	34	16	37	25	28	21
12	28	36	51	53	42	25
13	16	35	32	37	30	21
14	35	36	37	24	33	13
15	17	28	14	13	18	15
Total					445	253

For R-Chart:

$$
\begin{aligned}
UCL &= D_4\bar{R} = 2.28 \times 16.87 = 38.47 \\
CL &= \bar{R} = 16.87 \\
LCL &= D_3\bar{R} = 0 \times 16.87 = 0
\end{aligned}
$$

□

Problems

15.1 The following table gives the mean of 20 samples . Size of each sample is 25. The population standard deviation is 13. Draw the control chart for the mean:

31	25	31	27	20	18	17	29	30	32
19	23	28	37	25	31	27	33	20	27

15.2 In the production of certain rods, a process is said to be control if outside diameters have a mean = 2.5 inch, and a standard deviation of 0.002 inch. (i) Construct a control chart for the means of random sample of size 4. (ii) Means of such random samples taken at intervals were: 2.5014, 2.5022, 2.4995, 2.5076, 2.5040, 2.5001, 2.4993, 2.4962, 2.4966, 2.4971. Was the process ever out of control?

15.3 A drilling machine bores holes with a diameter of 0.50cm . and a standard deviation of 0.02cm.Construct a control chart for the mean of random samples of 4, showing the central line and the upper and lower control limits on a graph paper.

15.4 Measurements on average ($\bar{X}$) and ranges (R) from 20 samples each of 5 gave the following results:

$$\bar{\bar{X}} = 99.6,\ \bar{R} = 7.0$$

Determine the values of the control limits for drawing a Mean chart. (For $n = 5$, Mean Range = $2.32 \times$ Population standard deviation)

15.5 Samples of size 5 were taken for a study .The measurements noted are given below:

Sample No.	Observations				
1	42	65	75	78	87
2	19	24	80	81	81
3	36	54	69	77	84
4	51	74	75	78	132
5	61	78	94	109	136

Draw a Control Chart for mean if the factor for mean for n = 5 is 0.58. Only the central line and the upper and lower control limits may be drawn in the chart.

15.6 Range of 15 random samples of size 7 were:

20,34,10,18,20,16,21,9,11,14,22,19,22,20,11.

Construct a control chart for range (R-chart) and determine whether the process is within control or not ?(Give value of D_3 for size of sample 7 is 0.076 and of $D_4 = 1.924$)

15.7 In a factory 10 samples of 4 units each were taken . The range of samples are as follows: 5,6,5,7,9,4,8,6,4,6. Determine the central line and the control limits for R chart. (Give value of D_1 for size of sample 4 is 0 and $D_2 = 4.918, d_2 = 2.059$)

15.8 You are given the value of sample means ($\bar{X}$) and the ranges (R) for ten samples of size 5 each. Draw mean and range charts and comment on the state of control of the process:

Sample No.	1	2	3	4	5	6	7	8	9	10
$\bar{X}$	43	49	37	44	45	37	51	46	43	47
R	5	6	5	7	7	4	8	6	4	6

You may use the following control chart constants:

For $n = 5, A_2 = 0.58, D_3 = 0$ and $D_4 = 2.115$

15.9 Form a factory producing metal sheets, a sample of 5 sheets is taken every hour and the data obtained is as under. Draw a control chart for the mean and range and examine whether the process is under or not:

Sample No.	1	2	3	4	5	6	7	8	9	10
Sample Range	0.025	0.048	0.012	0.019	0.019	0.010	0.006	0.046	0.010	0.032
Mean thickness	0.025	0.032	0.042	0.022	0.028	0.010	0.025	0.040	0.026	0.029

You may use the following factors for finding out control limits:

N	A_2	D_3	D_4
10	0.58	0	2.115

15.10 Draw the control chart for $\bar{X}$ (mean) and $\bar{R}$(range) from the following data relating to 20 samples, each of size 5. you need not draw the charts.

Sample No.	$\overline{X}$	R_s	Sample No.	$\overline{X}$	R_s
1	38.3	15	11	32.6	31
2	33.8	11	12	22.8	12
3	24.4	12	13	21.6	29
4	36.6	24	14	28.8	22
5	27.4	18	15	28.8	16
6	30.6	33	16	24.4	19
7	31.2	21	17	30.4	20
8	27.4	20	18	25.4	34
9	24.0	29	19	37.8	19
10	29.4	18	20	31.4	17

15.2.2 Control Charts for Attributes

Control charts for attributes are used to measure quality characteristics that are counted rather than measured. Attributes are discrete in nature and entail simple yes-or-no decisions. For example, this could be the number of nonfunctioning light-bulbs, the number of rotten apples, or the number of complaints issued. Two of the most common types of control charts for attributes are p-charts and c-charts.

The qualitative characteristic of the manufactured product can not be expressed numerically; but simply we consider its presence and absence. Here we consider the number of defects per item. The control charts for the such situations are called *control charts for attributes.*

There are three types of control charts for attributes :

1. p-chart of proportion defectives
2. np-chart for number of defectives
3. c-chart for number of defects

15.2.2.1 p-Chart p-Charts are used to measure the proportion that is defective in a sample. The computation of the center line as well as the upper and lower control limits is similar to the computation for the other kinds of control charts. To obtain the p-chart we use following steps:

1. Find the proportion of defectives in each sample:

$$p = \frac{\text{Defective Units in a sample}}{\text{Size of sample}}$$

2. Find the mean ($\bar{p}$) of the sample proportions:
3. Compute control limits
   (a) Centre Line ,CL = $\bar{p} = \frac{p_1+p_2+p_3+......p_k}{k} = \frac{\sum p}{k}$
   (b) UCL $= \bar{p}+3\sqrt{\frac{\bar{p}\bar{q}}{n}}$
   (c) LCL $= \bar{p}-3\sqrt{\frac{\bar{p}\bar{q}}{n}}$

   where, $\bar{q} = 1-\bar{p}$

15.2.2.2 np-Chart np-Charts are used to measure the number of defective items in each sample. It is better to consider np-chart, when sample sizes are equal. For np-Charts the control limits are as follows:

$$CL = n\bar{p} = \frac{\text{Total numbers of defectives of all samples}}{\text{No of samples}}$$
$$UCL = n\bar{p}+3\sqrt{n\bar{p}\bar{q}}$$
$$LCL = n\bar{p}-3\sqrt{n\bar{p}\bar{q}}$$

Sometimes the lower control limit is negative, which occurs because the computation is an approximation of the binomial distribution. When this occurs, the LCL is rounded up to zero because we cannot have a negative control limit.

■ EXAMPLE 15.7

The following table gives the result of inspection of 20 samples of 100 items each taken on 20 working days . Draw a *p*-chart. What conclusion would you draw from the chart ?

Sample Number	Number of defective units	Sample Number	Number of defective units
1	6	11	10
2	2	12	4
3	4	13	6
4	11	14	11
5	20	15	22
6	6	16	8
7	10	17	0
8	19	18	3
9	4	19	23
10	21	20	10

Solution: Here, N=No of samples $\times$ Size of Sample = 20 $\times$ 100 =2000

$$\bar{p} = \frac{\Sigma p}{N} = \frac{200}{2000} = 0.10$$

$$\bar{q} = 1 - \bar{p} = 1 - 0.10 = 0.90$$

$$\therefore CL = \bar{p} = 0.10$$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}\bar{q}}{n}} = 0.10 + 3\sqrt{\frac{0.10 \times 0.90}{100}} = 0.19$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}\bar{q}}{n}} = 0.10 - 3\sqrt{\frac{0.10 \times 0.90}{100}} = 0.01$$

Sample Number	1	2	3	4	5	6	7	8	9	10
p	0.06	0.02	0.04	0.11	0.20	0.06	0.10	0.19	0.04	0.21
Sample Number	11	12	13	14	15	16	17	18	19	20
p	0.10	0.04	0.06	0.11	0.22	0.08	0.00	0.03	0.23	0.10

Draw the chart and observe that the process is not under control for sample number 5,10,15 and 19. □

■ **EXAMPLE 15.8**

From the following data construct an appropriate Control Chart:

Sample No.	Size of Sample	Defective Items
1	200	4
2	150	6
3	350	14
4	50	3
5	100	7
6	250	5
7	125	10
8	375	15
9	50	1
10	250	15

Solution: We can prepare p-chart or np-Chart:

p-chart.

$$\bar{p} = \frac{\sum p}{N} = \frac{80}{2000} = 0.04$$
$$\bar{q} = 1 - \bar{p} = 1 - 0.04 = 0.96$$

CL $= \bar{p} = 0.04$

UCL $= \bar{p} + 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$

Since the size of samples are not equal . we take n as the average of sample sizes :

$$n = \frac{N}{\text{No. of samples}} = \frac{2000}{10} = 200$$

$$\begin{aligned}
\therefore CL &= \bar{p} = 0.04 \\
UCL &= \bar{p} + 3\sqrt{\frac{\bar{p}\bar{q}}{n}} \\
&= 0.04 + 3\sqrt{\frac{0.04 \times 0.96}{200}} = 0.082 \\
LCL &= \bar{p} - 3\sqrt{\frac{\bar{p}\bar{q}}{n}} \\
&= 0.04 - 3\sqrt{\frac{0.04 \times 0.96}{200}} = -0.002 = 0
\end{aligned}$$

In this example the lower control limit is negative, the LCL is rounded up to zero because we cannot have a negative control limit.

np-chart.

$$\begin{aligned}
CL &= n\bar{p} = 200 \times 0.04 = 8 \\
UCL &= n\bar{p} + 3\sqrt{n\bar{p}\bar{q}} \\
&= 8 + 3\sqrt{200 \times 0.04 \times 0.96} = 16.31 \\
LCL &= n\bar{p} - 3\sqrt{200 \times 0.04 \times 0.96} = -0.31 = 0
\end{aligned}$$

□

■ EXAMPLE 15.9

20 samples were taken, each of 100 units. The number of defectives found are as under. Draw a np-chart for number of defectives:

2	6	2	4	4	18	0	4	10	18
2	4	6	4	8	0	2	2	4	0

Solution: Here N=No of samples × Size of Sample = 20 × 100 =2000

$$\bar{p} = \frac{\sum p}{N}$$
$$= \frac{100}{2000} = 0.05$$
$$\bar{q} = 1 - \bar{p} = 1 - 0.05 = 0.95$$

$$CL = n.\bar{p} = 100 \times 0.05 = 5$$
$$UCL = n\bar{p} + 3\sqrt{n.\bar{p}.\bar{q}} = 5 + (3 \times \sqrt{100 \times 0.05 \times 0.95}) = 11.54$$
$$LCL = n\bar{p} - 3\sqrt{n.\bar{p}.\bar{q}} = 5 - (3 \times \sqrt{100 \times 0.05 \times 0.95}) = -1.54 = 0 \text{ (Negative not allowed.)}$$

□

15.2.2.3 c-Chart When it is necessary to control the number of defects , then we construct a c-chart . For example, number spots on a paper, number of defects on a polished plate, number of defects in a piece of cloth, etc. The process of constructing c-chart is as follows :

1. Central Line : First of all calculate the average number of defects ($\bar{c}$) where $\bar{c}$ is given by

$$\text{Central Line} = \bar{c} = \frac{\text{Total number of defects in all the observed items}}{\text{No. of items}}$$

2. Control Limits : Control limits based on Poisson distribution are as follows:
   (a) Upper Control Limit UCL $= \bar{c} + 3\sqrt{\bar{c}}$
   (b) Lower Control Limit LCL $= \bar{c} - 3\sqrt{\bar{c}}$

■ EXAMPLE 15.10

The number of defects found in each one of the 15 pieces of 2×2 metre of a synthetic fiber cloth is given below. Construct an appropriate control chart and state your conclusion. Also show three lines on the graph paper:
3,7,12,5,21,4,3,20,0,8,10,20,9,7,6.

Solution: Since the number of defects are given , so the suitable chart is c-chart.

$$CL = \bar{c} = \frac{\text{Total number of defects in all the observed items}}{\text{No. of items}}$$
$$= \frac{139}{15}$$
$$UCL = \bar{c} + 3\sqrt{\bar{c}} = 9 + 3\sqrt{9} = 18$$
$$LCL = \bar{c} - 3\sqrt{\bar{c}} = 9 - 3\sqrt{9} = 0$$

Since three points are outside the control limits, so the process is out of control. □

■ **EXAMPLE 15.11**

The following table gives the number of defects found in each of the 120 carpets manufactured in a factory. On the basis of these data, construct a chart for the number of defects per carpet:

No.of defects(x)	1	4	5	8
No. of carpets having x defects (f)	37	35	27	21

Solution: Construction of c-chart.

No. of Defects (x)	No.of Carpets having x defects (f)	fx
1	37	37
4	35	140
5	27	135
8	21	168
	120	$\sum fx = 480$ =Total defects

$$CL = \bar{c} = \frac{480}{120} = 4$$
$$= \bar{c} + 3\sqrt{\bar{c}} = 4 + 3\sqrt{4} = 10$$
$$= \bar{c} - 3\sqrt{\bar{c}} = 4 - 3\sqrt{4} = -2 \text{ or } 0$$

□

Problems

15.11 A manufactured finds that on an average 2 in 100 of the items produced by him is defective. Draw a Control chart for number of defective items for sample size 100.

15.12 The following table gives the result of inspection of 20 samples of 100 items each taken as 20 working days. Draw a *p*-chart. What conclusions would you draw from the chart.

Sample No.	No. Inspected	No. defective	Sample No.	No. Inspected	No. Defective
1	100	9	11	100	10
2	100	17	12	100	6
3	100	8	13	100	7
4	100	7	14	100	18
5	100	12	15	100	16
6	100	5	16	100	10
7	100	11	17	100	5
8	100	16	18	100	14
9	100	14	19	100	7
10	100	15	20	100	13

15.13 The following data refer to visual defects found at inspection of the first 10 sample of size 100 each. Use the data to obtain the central line and upper and lower control limits for percentage defectives:

Sample No. :	1	2	3	4	5	6	7	8	9	10
No. of defectives :	2	1	1	2	3	2	4	0	3	2

15.14 The following table given the result of inspection of 20 samples of 100 items each taken on 20 working days draw a *np*-chart. What conclusion would you draw from the chart:

sample no.	1	2	3	4	5	6	7	8	9	10
No.of defectives	0	2	4	6	6	4	0	2	4	8
sample No.	11	12	13	14	15	16	17	18	19	20
No.of defectives	8	0	4	6	14	0	2	2	6	2

15.15 A manufacturer finds that on average 1 in 10 of the items produced by him is defective A. Few days later he finds 20 items in a sample of 100 items defective. Is the process out of control? Give reasons for your answer.

15.16 The total number of defects in 30 large size samples at a work station was 480. Apply the Poisson distribution to determine the central line and the upper and lower control limits for the number of defects in a sample. Draw the three lines of the control chart on graph paper.

15.17 The following table gives the result of inspection of pieces of woolen suiting cloth:

Piece No.	1	2	3	4	5	6	7	8	9	10
No of defects	4	3	6	3	0	1	3	5	7	8

Construct an appropriate control chart and comment whether the process is in a state of control or not.

15.18 18 items were observed closely and the numbers of defects in their texture noted as below supposing that the number of defects following the Poisson distribution draw a control chart for the central line and upper and lower control limits for the number of defects:

No. of Defects	0	1	2	3	4	5	6
No.of carpets	0	1	2	4	3	5	3

UNIT 4
NUMERICAL TECHNIQUES - I

CHAPTER 16

ROOT FINDING

Calculating the root of an equation

$$f(x) = 0 \tag{16.1}$$

is a widely used problem in engineering and applied mathematics. In this chapter we will explore some simple numerical methods for solving this equation.

The function $f(x)$ will usually have at least one continuous derivative, and often we will have some estimate of the root that is being computed. By using this information, most of the numerical methods compute a sequence of increasingly improved roots for 16.1. These methods are called iteration methods.

16.1 RATE OF CONVERGENCE

Here, the speed at which a convergent sequence approaches its limit is called the rate of convergence. This concept is of practical importance if we deal with a sequence of successive approximations for an iterative method, as then typically fewer iterations are needed to yield a useful approximation if the rate of convergence is higher.

16.1.1 Convergence of A Sequence

A sequence $< x_n >$ is said to be converges to α if for every $\varepsilon > 0$ there is an integer $m > 0$ such that if $n > m$ then $|x_n - \alpha| < \varepsilon$. The number α is called the limit of the sequence and we sometimes write $x_n \to c$.

16.1.2 Convergence speed for iterative methods

Let $< x_n >$ be sequence of successive approximations of a root $x = \alpha$ of the equation $f(x) = 0$. Then the sequence $< x_n >$ is said to be converges to α with order q if

$$\lim_{n\to\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^q} = \mu \tag{16.2}$$

where $\mu > 0$. Constant μ is called the rate of convergence.

Particularly, if $|x_{n+1} - \alpha| = \mu |x_n - \alpha|,\ n \geq 0,\ \ 0 < \mu < 1$, then convergence is called linear or of first order. Convergence with order 2 is called quadratic convergence, and convergence with order 3 is called cubic convergence.

16.2 BISECTION METHOD

Suppose $f(x)$ is continuous on an interval $[a,b]$, such that

$$f(a).f(b) < 0 \tag{16.3}$$

Then $f(x)$ changes sign on $[a,b]$, and $f(x) = 0$ has at least one root on the interval. Bisection method repeatedly halve the interval $[a,b]$, keeping the half on which $f(x)$ changes sign. It guaranteed to converge to a root.

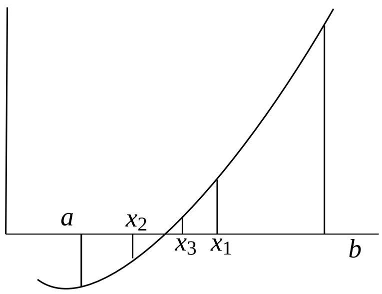


Figure 16.1. Bisection Method

More preciously, Suppose that we are given an interval $[a,b]$ satisfying 16.3 and an error tolerance $\varepsilon > 0$. Then the bisection method is consists of the following steps:

[B1.] Compute $c = (a+b)/2$

[B2.] If $b - c \leq \varepsilon$, then accept c as the root and stop the procedure.

[B3.] If $f(a).f(c) \leq 0$, then set $b = c$ else, set $a = c$. Go to step B1.

■ **EXAMPLE 16.1**

Find one of the real root of $f(x) \equiv x^6 - x - 1 = 0$ accurate to within $\varepsilon = 0.001$.

Solution: We consider the interval $[1,2]$, choose $a = 1,\ b = 2$; then $f(a) = -1,\ f(b) = 61$. This shows that condition 16.3 is satified for this interval. Now we may apply algorithm B1 to B3

Iterations:1

$$c = \frac{a+b}{2} = \frac{1+2}{2} = 1.5$$

$$b - c = 0.5 > \varepsilon$$

Move to next step, decide new interval

$$f(a).f(c) = (-1).(8.8906) = -8.8906 < 0$$

This shows new interval $[a, c]$ is valid interval. i.e. replace $b = c$. Now $a = 1$ and $b = 1.5$

Table 16.1.

n	a	b	c	$b-c$	$f(c)$
1	1.0000	2.0000	1.5000	0.5000	8.8906
2	1.0000	1.5000	1.2500	0.2500	1.5647
3	1.0000	1.2500	1.1250	0.1250	-0.0977
4	1.1250	1.2500	1.1875	0.0625	0.6167
5	1.1250	1.1875	1.1563	0.0313	0.2333
6	1.1250	1.1563	1.1406	0.0156	0.0616
7	1.1250	1.1406	1.1328	0.0078	-0.0196
8	1.1328	1.1406	1.1367	0.0039	0.0206
9	1.1328	1.1367	1.1348	0.0020	0.0004
10	1.1328	1.1348	1.1338	0.00098	-0.0096

Iterations:2

$$c = \frac{a+b}{2} = \frac{1+1.5}{2} = 1.25$$

$$b - c = 0.25 > \varepsilon$$

Move to next step, decide new interval

$$f(a).f(c) = (-1).(1.5647) = -1.5647 < 0$$

This shows new interval $[a, c]$ again, is a valid interval. i.e. replace $b = c$. Now $a = 1$ and $b = 1.25$

Repeat this process. Further steps are given in table 16.1.

Note after 10th iteration $b - c < \varepsilon$. Stop iteration process at this step. Hence final approximated root is 1.1338. □

■ EXAMPLE 16.2

Find a real root of the equation $x^2 - 2x - 5 = 0$.

Solution: Let $f(x) = x^3 - 2x - 5$. Then $f(2) = -1$ and $f(3) = 16$, therefore a root lies between 2 and 3 and we take

$$c = \frac{2+3}{2} = 2.5$$

Since $f(a).f(c) < 0$, we choose $[2, 2.5]$ as the new interval. Then, again

$$c = \frac{2+2.5}{2} = 2.25$$

and $f(a).f(c) < 0$. Proceeding in this way, the following table is obtained.

Table 16.2.

n	a	b	x	f(x)
1	2.00000	3.00000	2.50000	5.6250
2	2.00000	2.50000	2.25000	1.8906
3	2.00000	2.25000	2.12500	0.3457
4	2.00000	2.12500	2.06250	-0.3513
5	2.06250	2.12500	2.09375	-0.0089
6	2.09375	2.12500	2.10938	0.1668
7	2.09375	2.10938	2.10156	0.0786
8	2.09375	2.10156	2.09766	0.0347
9	2.09375	2.09766	2.09570	0.0129
10	2.09375	2.09570	2.09473	0.0019
11	2.09375	2.09473	2.09424	-0.0035
12	2.09424	2.09473		

At $n = 12$, it is seen that the difference between two successive iterates is 0.0005 . Therefore root is 2.094 □

16.2.1 Convergence of Bisection Method

In Bisection Method, the original interval is divided into half in each iteration. Thus, if error in ith iteration is $\varepsilon_i = |x_i - \alpha|$, and error in $(i+1)$th iteration is $\varepsilon_{i+1} = |x_{i+1} - \alpha|$, then we have

$$\frac{\varepsilon_{i+1}}{\varepsilon_i} = \frac{1}{2}$$

Then by definition in section 16.1, sequence of approximations in bisection method is linearly convergent. Sometimes it is said Bisection method is linearly convergent. Further the rate of convergence of this method is 0.5.

Problems

16.1 Transcendental equation is given as

$$f(x) = 2^x - x - 3$$

Calculate $f(x)$ for $x = -4, -3, -3, 1, 0, 1, 2, 3, 4$ and compute between which integers values roots are lying.

16.2 Use the bisection method to find the indicated roots of the following equations. Use an error tolerance $\varepsilon = 0.0001$.

- **a)** The real root of $x^3 - x^2 - x - 1 = 0$.
- **b)** The root of $x = 1 + 0.3\cos(x)$.
- **c)** The root of $x = e^{-x}$.
- **d)** The smallest positive root of $e^{-x} = \sin(x)$.
- **e)** The real root of $e^x = 3x$.

f) The root of the equation $3x - \sqrt{1+\sin x} = 0$

16.3 Find the real root correct to three decimal places for the following equations:

a) $x^3 - x - 4 = 0$

b) $\log x = \cos x$

c) $x^3 - x^2 - 1 = 0$

d) $x^3 + x^2 - 1 = 0$

16.4 Find the real root of the equation $x \log x = 1.2$ by Bisection Method correct up to four decimal places. Also implement a C-Program for the same.

16.5 Find a positive real root of $x - \cos x = 0$ by Bisection Method, correct up to 3 decimal places between 0 and 1.

16.6 Use bisection Method to find out the positive root of 30 correct upto 4 decimal places.

16.7 The equation $x^2 - 2x - 3\cos x = 0$ is given. Locate the smallest root in magnitude in an interval of length one unit.

16.8 Find the smallest positive root of $x^3 - 9x + 1 = 0$,using Bisection method correct to three decimal places.

16.9 Find the root of $\tan x + x = 0$ up to two decimal places which lies between 2 and 2.1 using Bisection method.

16.10 Find a positive root of the equation $xe^x = 1$ which lies between 0 and 1.

16.11 Write computer program in a language of your choice which implements bisection method to compute real root of the equation $3x + \sin x - e^x = 0$ in a given interval.

16.12 The smallest positive root of the equation $f(x) = x^4 - 3x^2 + x - 10 = 0$ is to be obtained.

a) Find an interval of unit length which contains this root.

b) Perform two iterations of the bisection method.

16.3 REGULA FALSI METHOD

This method is also called as false position method. Consider the equation 16.1. Let a and b, such that $a < b$, be two values of x such that $f(a)$ and $f(b)$ are of opposite signs. Then the graph of $y = f(x)$ crosses the x-axis at some point between a and b.

The equation of the chord joining the two point $(a, f(a))$ and $(b, f(b))$ is

$$y - f(a) = \frac{f(b) - f(a)}{b - a}(x - a) \tag{16.4}$$

On x axis $y = 0$, hence

$$0 - f(a) = \frac{f(b) - f(a)}{b - a}(x - a)$$

or

$$x = \frac{af(b) - bf(a)}{f(b) - f(a)} = x_1(say)$$

Now if function is considered as a straight line, the intersection of chord produces as approximate root value. Further very similar to Bisection Method root lies in between a and x_1 or between x_1 and b, depends upon the fact $f(a)f(x_1) < 0$ or $f(x_1)f(b) < 0$ respectively. Thus we may concentrate on smaller interval in which root lies. We repeat this process with interval in which root lies.

Suppose that we are given an interval $[a, b]$ satisfying 16.3 and an error tolerance $\varepsilon > 0$. Then the Regula Falsi Method is consists of the following steps:

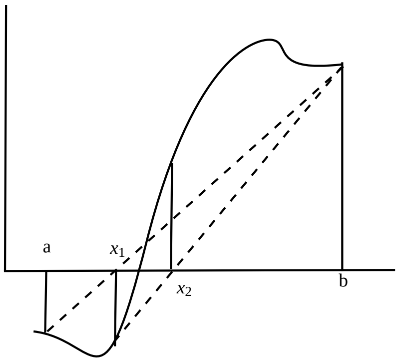


Figure 16.2. Regula Falsi Method

R1. Compute $c = \dfrac{af(b) - bf(a)}{f(b) - f(a)}$

R2. If the difference with two consecutive c is less than or equal to ε, then accept c as the root and stop the procedure.

R3. If $f(a).f(c) \leq 0$, then set $b = c$ else, set $a = c$. Go to step B1.

■ EXAMPLE 16.3

Find the root of the equation

$$xe^x = \cos x \tag{16.5}$$

in the interval [0,1] using Regula Falsi method correct to four decimal places.

Solution: Let $f(x) = \cos x - xe^x$ Here $f(0) = 1$ and $f(1) = -2.17798$. i.e. the root lies in (0,1). Now we may apply algorithm R1 to R3. To compute c, we used

$$c = \frac{af(b) - bf(a)}{f(b) - f(a)}$$

Iterations:1

$$c = \frac{af(b) - bf(a)}{f(b) - f(a)} = \frac{0.(-2.17798) - 1.(1)}{-2.17798 - 1} = 0.31467$$

Here,

$$f(a).f(c) = f(0)f(0.31467) = (1)(0.51987) > 0$$

This shows that the root lies in between 0.31467 and 1.

Iterations:2 Now taking $a = 0.31467$ and $b = 1$

$$c = \frac{af(b) - bf(a)}{f(b) - f(a)} = \frac{(0.31467)(1) - (1)(0.51987)}{1 - 0.51987} = 0.44673$$

Here $f(c) = 0.20356$ Move to next step, decide new interval

$$f(a).f(c) = (0.51987).(0.20356) > 0$$

This shows new interval $[c, b]$ again, is a valid interval. i.e. now $a = 0.44673$, adn $b = 1$ for the next iteration. All the next iterations are given in the table:

The values are computed in table 16.3.. Since the difference of two consecutive approximation is

Table 16.3. Regula Falsi Method Example 1

n	a	b	c	$f(c)$
1	0.00000	1.00000	0.31467	0.51987
2	0.31467	1.00000	0.44673	0.20354
3	0.44673	1.00000	0.49402	0.07080
4	0.49402	1.00000	0.50995	0.02361
5	0.50995	1.00000	0.51520	0.00776
6	0.51520	1.00000	0.51692	0.00254
7	0.51692	1.00000	0.51748	0.00083
8	0.51748	1.00000	0.51767	0.00027
9	0.51767	1.00000	**0.51773**	0.00009
10	0.51773	1.00000	**0.51775**	0.00003

less than required error accepted. Hence root is 0.51775. □

Problems

16.13 Find a real root of the following equations correct to four decimal places using the method of false position (Regula Falsi Method).

a) $x^3 - 2x - 5 = 0$
b) $x \log x = 1.2$
c) $x^4 - x^3 - 2x^2 - 6x - 4 = 0$
d) $x^3 - 5x + 3 = 0$

16.14 Find the root of the equation $\tan x + \tanh x = 0$ which lies in the interval $(1.6, 3.0)$ correct to three significant digits using method of false position.

16.15 Find real cube root of 18 by Regula-falsi method.

16.16 Determine the real roots $f(x) = x^3 - 98$ using False position method within $E - s = 0.1\%$.

16.17 Solve the following equations by Regula-Falsi method.

a) $(5 - x)e^x = 5$ near $x = 5$.
b) $2x - \log x = 7$ lying between 3.5 and 4.
c) $x^3 + x^2 - 3x - 3 = 0$ lying between 1 and 2.
d) $x^4 + x^3 + -7x^2 - x + 5 = 0$ lying between 2 and 43.

16.18 Illustrate false position method by plotting the function on graph and discuss the speed of convergence to the root. Develop the algorithm for computing the roots using false-position technique.

16.19 Find all the roots of $\cos x - x^2 - x = 0$ to 5 decimal places.

16.4 NEWTON RAPHSON METHOD

When an approximate value of a root of an equation is given, this method is used to obtain better and closer approximation to the root. Let x_0 be an approximation of a root of the given equation $f(x)=0$, Let $x_1=x_0+h$ be the exact approximation of the root. Then $f(x_0+h)=0$.

By Taylor's theorem, we have

$$f(x_0+h) = f(x_0)+hf'(x_0)+\frac{h^2}{2!}f''(x_0)+\ldots=0$$

Since h is small, we can neglect second, third and higher degree terms in h and thus we get

$$f(x_0)+hf'(x_0)=0$$

or

$$h=-\frac{f(x_0)}{f'(x_0)};\quad f'(x_0)\neq 0$$

Hence

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

We may iterate the process to refine the root. In general, we may write

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)} \tag{16.6}$$

This result is known as Newton-Raphson formula.

■ EXAMPLE 16.4

Solve $\sin x = 1+x^3$ using Newton-Raphson Method.

Solution: Let $f(x)=\sin x-1-x^3$, then $f'(x)=\cos x-3x^2$. Then Newton-Raphson formula for this problem reduces to

$$x_{n+1}=x_n-\frac{\sin x_n-1-x_n^3}{\cos x_n-3x_n^2}$$

i.e.

$$x_{n+1}=\frac{x_n\cos x_n-\sin x_n-2x_n^3+1}{\cos x_n-3x_n^2} \tag{16.7}$$

Now, since $f(-1)<0$ and $f(-2)>0$, root lies in between -1 and -2. Let $x_0=-1.1$ be the initial approximation. then successive iteration from 16.7 are

$$x_1=\frac{x_0\cos x_0-\sin x_0-2x_0^3+1}{\cos x_0-3x_0^2}=\frac{4.05425}{-3.17640}=-1.27636$$

Similarly, $x_2=-1.249746$, $x_3=-1.2490526$, $x_4=-1.2490522$. In x_2 and x_3 6 decimal places are same. i.e. approximated root up to six decimal place is -1.249052. □

■ EXAMPLE 16.5

Find a root of the equation $x\sin x+\cos x=0$

Solution: We have $f(x)=x\sin x+\cos x$, and $f'(x)=x\cos x$, therefore the iteration formula is

$$x_{n+1}=x_n-\frac{x_n\sin x_n+\cos x_n}{x_n\cos x_n}$$

with $x_0 = \pi$, the successive iterations are given in the following table

n	x_n	$f(x_n)$	x_{n+1}
0	3.1416	-1.0000	2.8233
1	2.8233	-0.0662	2.7986
2	2.7986	-0.0006	2.7984
3	2.7984	0.0000	2.7984

□

16.4.1 Convergence of Newton-Raphson Method

In this section, we will see the condition for the convergence of Newton-Raphson method. We have

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

This is an iteration[1] method where

$$x_{n+1} = \phi(x_n); \text{ and } \phi(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}$$

In general,

$$x = \phi(x); \text{ and } \phi(x) = x - \frac{f(x)}{f'(x)}$$

As iteration method converges for $|\phi'(x)| < 1$, that is,

$$\left| \frac{d}{dx}\left(x - \frac{f(x)}{f'(x)} \right) \right| < 1$$

or

$$\left| 1 - \frac{[f'(x)]^2 - f(x) f''(x)}{[f'(x)]^2} \right| < 1$$

or

$$\left| \frac{f(x) f''(x)}{[f'(x)]^2} \right| < 1$$

i.e.

$$\left| f(x) f''(x) \right| < [f'(x)]^2$$

The interval containing the root α of $f(x) = 0$ should be selected in which the above is satisfied.

16.4.2 Rate Convergence of Newton-Raphson Method

Let x_n and x_{n+1} be two successive approximations to the actual root α of $f(x) = 0$. Then

$$x_n - \alpha = \varepsilon_n$$

and

$$x_{n+1} - \alpha = \varepsilon_{n+1}$$

[1]**Fixed Point Iteration Method**: In this method equation $f(x) = 0$ is written in the form of $x = \phi(x)$, such that $|\phi'(x)| < 1$. Then sequence generated by $x_{n+1} = \phi(x_n)$ converges to the root of the equation $f(x) = 0$.

where e_i is the error in ith iteration. Now equation 16.6 gives,

$$\alpha+\varepsilon_{n+1}=\alpha+\varepsilon_n-\frac{f(\alpha+\varepsilon_n)}{\alpha+\varepsilon_n}$$
$$\varepsilon_{n+1}=\varepsilon_n-\frac{f(\alpha+\varepsilon_n)}{\alpha+\varepsilon_n}$$

By Taylor's Theorem[2]

$$=\varepsilon_n-\frac{f(\alpha)+\varepsilon_n f'(\alpha)+\frac{1}{2}(\varepsilon_n^2 f''(\alpha))+\ldots}{f'(\alpha)+\varepsilon_n f''(\alpha)+\frac{1}{2}(\varepsilon_n^2 f'''(\alpha))+\ldots}$$
$$=\varepsilon_n-\frac{\varepsilon_n f'(\alpha)+\frac{1}{2}(\varepsilon_n^2 f''(\alpha))}{f'(\alpha)+\varepsilon_n f''(\alpha)}\ (As\ f(\alpha)=0)$$
$$=\frac{1}{2}\varepsilon_n^2\frac{f''(\alpha)}{f'(\alpha)+\varepsilon_n f''(\alpha)}$$
$$=\frac{1}{2}\frac{\varepsilon_n^2 f''(\alpha)}{f'(\alpha)}\ \text{Omitting derivatve of order higher than two}$$

Thus

$$\frac{\varepsilon_{n+1}}{\varepsilon_n^2}=\frac{f''(\alpha)}{2f'(\alpha)}=k\ \ (say)$$

which shows that at any stage, the subsequent error is proportional to the square of the previous error. Hence, Newton-Raphson method has a quadratic convergence. In other words, its order of convergence is 2. Again rate of convergence is k.

Problems

16.20 Find an iterative formula to find $\sqrt{m}$, where m is a positive number and hence, find $\sqrt{12}$ correct to four decimal places.

16.21 Find the roots of the following using Newton-Raphson formula.

a) $x^3+2x^2+10x-20=0$
b) $x\log x=1.2$
c) $\sin x=1-x$
d) $x^3-5x+3=0$
e) $x-\cos x=0$
f) $\sin x=\frac{x}{2}$
g) $x+\log x=2$
h) $\tan x=x$
i) $\log x=\cos x$
j) $x^2+4\sin x=0$

16.22 By using Newton-Raphson's method, find the root of $x^4-x-10=0$ which is near to $x=2$ correct to three places of decimal.

[2]**Taylor's Theorm:**

$$f(x+h)=f(x)+hf'(x)+\frac{h^2}{2!}f''(x)+\ldots+\frac{h^n}{n!}f^{(n)}(x)+\ldots$$

16.23 Compute one positive root of $2x - \log_{10} x = 7$ by Newton-Raphson method correct to four decimal places.

16.24 Develop an algorithm using Newton-Raphson method to find the fourth root of a positive number N and hence find $\sqrt[4]{32}$.

16.25 Prove the recurrence formula

$$x_i + 1 = \frac{1}{3}\left[2x_i + \frac{N}{x_i^2}\right]$$

for finding the cube root of N. Hence find cube root of 63.

16.26 Show that the iterative formula for finding the reciprocal of n is $x_{i+1} = x_i(2 - nx_i)$, hence find the value of $\dfrac{1}{31}$.

16.27 Determine p, q and r so that the order of the iterative method

$$x_{n+1} = px_n + \frac{qa}{x_n^2} + \frac{ra^2}{x_n^5}$$

for $a\frac{1}{3}$ becomes as high as possible [Hint: $p+q+r=1, p-2q-5r=0, 3q+15r=0$]

16.28 Find all positive roots of the equation $10\int e^{-x^2}dt - 1 = 0$

16.29 The equation $2e^{-x} = \frac{1}{x+2} + \frac{1}{x+1}$ has two roots greater than -1. Calculate these roots correct to five decimal place.

16.30 The equation $x = 0.2 + 0.4\sin\dfrac{x}{b}$ where b is a parameter, has one solution near $x = 0.3$, the parameter is known only with some uncertainty: $b = 1.2 \pm 0.05$. Calculate the root with an accuracy reasonable with respect to the uncertainty of b.

16.31 Find the positive root the equation

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}e^{0.3x}$$

correct to 6 decimal place.

16.32 Show that the equation

$$f(x) = cos\frac{\pi(x+1)}{8} + 0.148x - 0.9062 = 0$$

has one root in the intervals $(-1, 0)$ and one in $(0, 1)$. Calculate the negative root correct to 4 decimal.

CHAPTER 17

INTERPOLATION

Interpolation is the process of finding a function whose graph passes through a set of given points. It is a method of constructing new data points within the range of a discrete set of known data points.

In engineering and science one often has a number of data points, as obtained by sampling or experimentation, and tries to construct a function which closely fits those data points. This is called curve fitting or regression analysis. Interpolation is a specific case of curve fitting, in which the function must go exactly through the data points.

17.1 FINITE DIFFERENCE: NOTIONS

Let $y = f(x)$ be a discrete function. Consider the given data set of $n+1$ values (x_0, y_0), (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), where x differ by a quantity h, i.e. values of x are equidistant with interval distance h. The value of x is usually called *argument* and the corresponding function value y is called *entry*. In following subsection, we discuss three types of finite differences:

1. Forward Differences
2. Backward Differences
3. Central Differences

17.1.1 Forward Differences

For the given data set of $n+1$ values (x_0, y_0), (x_1, y_1),(x_2, y_2),..., (x_n, y_n), the quantities $y_1 - y_0$, $y_2 - y_1$, ..., $y_n - y_{n-1}$ are called *differences*, particularly *first differences*, and are denoted by Δy_0,

Δy_1, ..., Δy_{n-1} respectively. In general, first forward differences are given by

$$\Delta y_n = y_{n+1} - y_n$$

The symbol Δ is called *forward difference operator*. Further second forward differences are defined as the differences of the first differences. i.e.,

$$\Delta^2 y_n = \Delta(\Delta y_n) = \Delta(y_{n+1} - y_n)$$

That is,

$$\Delta^2 y_n = \Delta y_{n+1} - \Delta y_n$$

Here, Δ^2 is called second forward difference operator. Similarly, other higher order forward difference may be computed. In general,

$$\Delta^{m+1} y_n = \Delta^m y_{n+1} - \Delta^m y_n$$

Table 17.1. Forward Difference Table

x	y	Δ	Δ^2	Δ^3
x_0	y_0			
		$\Delta y_0 = y_1 - y_0$		
x_1	y_1		$\Delta^2 y_0 = \Delta y_1 - \Delta y_0$	
		$\Delta y_1 = y_2 - y_1$		$\Delta^3 y_0 = \Delta^2 y_1 - \Delta^2 y_0$
x_2	y_2		$\Delta^2 y_1 = \Delta y_2 - \Delta y_1$	
		$\Delta y_2 = y_3 - y_2$		
x_3	y_3			

17.1.2 Backward Differences

The differences $y_1 - y_0$, $y_2 - y_1$, ..., $y_n - y_{n-1}$ when denoted by ∇y_1, ∇y_2, ..., ∇y_n respectively, are called the first backward differences. Here, ∇ (read as Nabla) is *backward differences operator*. Thus we have,

$$\nabla y_r = y_r - y_{r-1}$$

$$\nabla^2 y_r = \nabla y_r - \nabla y_{r-1}$$

That is, in general,

$$\nabla^{m+1} y_r = \nabla^m y_r - \nabla^m y_{r-1}$$

17.1.3 Central Differences

The differences $y_1 - y_0$, $y_2 - y_1$, ..., $y_n - y_{n-1}$ when denoted by $\delta y_{1/2}$, $\delta y_{3/2}$, ..., $\delta y_{(2n-1)/2}$ respectively, are called the central differences. Here, δ is called as central difference operator. Thus we have,

$$\delta^{m+1} y_{(2r-1)/2} = \delta^m y_r - \delta^m y_{r-1}$$

Table 17.2. Backward Difference Table

x	y	∇	∇^2	∇^3
x_0	y_0			
		$\nabla y_1 = y_1 - y_0$		
x_1	y_1		$\nabla^2 y_2 = \nabla y_2 - \nabla y_1$	
		$\nabla y_2 = y_2 - y_1$		$\nabla^3 y_3 = \nabla^2 y_3 - \nabla^2 y_2$
x_2	y_2		$\nabla^2 y_3 = \nabla y_3 - \nabla y_2$	
		$\nabla y_3 = y_3 - y_2$		
x_3	y_3			

Table 17.3. Central Difference Table

x	y	δ	δ^2	δ^3
x_0	y_0			
		$\delta y_{1/2} = y_1 - y_0$		
x_1	y_1		$\delta^2 y_1 = \delta y_{3/2} - \delta y_{1/2}$	
		$\delta y_{3/2} = y_2 - y_1$		$\delta^3 y_{3/2} = \delta^2 y_2 - \delta^2 y_1$
x_2	y_2		$\delta^2 y_2 = \delta y_{5/2} - \delta y_{3/2}$	
		$\delta y_{5/2} = y_3 - y_2$		
x_3	y_3			

17.2 SOME OTHER DIFFERENCE OPERATORS

So far we have studied the operators Δ, ∇ and δ. Here we will discuss some more operators which play a vital role in numerical analysis.

17.2.1 Shifting Operator

If h is the interval of differencing in the argument x then the operator E is defined as

$$E f(x) = f(x+h)$$

and

$$E^{-1} f(x) = f(x-h)$$

Sometimes it is also called translation operator due to it results value of the function for the next argument. Furhter we observe that

$$E^2 f(x) = E(E f(x)) = E f(x+h) = f(x+2h)$$

In general, for all integral values of n

$$E^n f(x) = f(x+nh)$$

17.2.2 The D Operator

The differential coefficient of y with respect to x is denoted by Dy, where $D \equiv \dfrac{d}{dx}$. Here D is called as differential operator or simply operator D. We may denote nth derivative of y with respect to x as $D^n y$. Here D^n is called nth differential operator.

17.2.3 The Mean Operator

The mean operator is denoted by μ and is defined as

$$\mu[y_x] = \frac{1}{2}\left[y_{x+\frac{h}{2}} + y_{x-\frac{h}{2}}\right]$$

17.3 SOME IMPORTANT RELATIONS

The shift operator E is fundamental operator. All other derivatives may be derive from it. In this section we discuss some representation of other operators in terms of E.

We have,

$$\begin{aligned}\Delta y_x &= y_{x+h} - y_x = Ey_x - y_x \\ &= (E-1)y_x\end{aligned}$$

Which implies $\Delta \equiv E - 1$, i.e.

$$E \equiv 1 + \Delta \tag{17.1}$$

Similarly for the backward operator, we have

$$\begin{aligned}\nabla y_x &= y_x - y_{x-h} = y_x - E^{-1}y_x \\ &= (1-E^{-1})y_x\end{aligned}$$

That is,

$$\nabla \equiv 1 - E^{-1} \tag{17.2}$$

$$\begin{aligned}Ey_x &= y_{x+h} \\ &= y_x + hDy_x + \frac{h^2}{2!}D^2y_x + \ldots \text{ (Taylor's Theorem)} \\ &= \left[1 + hD + \frac{h^2D^2}{2!} + \ldots\right]y_x \\ &= e^{hD}y_x\end{aligned}$$

That is,

$$E \equiv e^{hD} \tag{17.3}$$

By using relation 17.1 and rearranging, we get

$$\Delta \equiv e^{hD} - 1 \tag{17.4}$$

■ EXAMPLE 17.1

Using the method of separation of symbols, show that

$$\Delta^n u_{x-n} = u_x - nu_{x-1} + \frac{n(n-1)}{2}u_{x-2} + \ldots + (-1)^n u_{x-n}$$

Solution:

$$u_x - nu_{x-1} + \frac{n(n-1)}{2}u_{x-2} + \cdots + (-1)^n u_{x-n}$$

$$\begin{aligned}
&= u_x - nE^{-1}u_x + \frac{n(n-1)}{2}E^{-2}u_x + \ldots + (-1)^n E^{-n}u_x \\
&= \left[1 - nE^{-1} + \frac{n(n-1)}{2}E^2 + \ldots + (-1)^n E^{-n}\right] u_x \\
&= (1-E^{-1})^n u_x \\
&= \left(\frac{E-1}{E}\right)^n u_x = \frac{\Delta^n}{E^n}u_x = \Delta^n E^{-n} u_x \\
&= \Delta^n u_{x-n}
\end{aligned}$$

□

■ EXAMPLE 17.2

Given $u_0 = 3$, $u_1 = 12$, $u_2 = 81$, $u_3 = 200$, $u_4 = 100$ and $u_5 = 8$. find the value od $\Delta^5 u_0$

Solution: We have

$$\begin{aligned}
\Delta^5 u_0 &= (E-1)^5 u_0 \\
&= (E^5 - 5E^4 + 10E^3 - 10E^3 + 5E - 1)u_0 \\
&= (E^5 u_0 - 5E^4 u_0 + 10E^3 u_0 - 10E^2 u_0 + 5Eu_0 - 1u_0) \\
&= u_5 - 5u_4 + 10u_3 - 10u_2 + 5u_1 - u_0 \\
&= 8 - 5 \times 100 + 10 \times 200 - 10 \times 81 + 5 \times 12 - 3 \\
&= 755
\end{aligned}$$

□

■ EXAMPLE 17.3

Evaluate $\Delta^n \sin(ax+b)$, given $h = 1$

Solution:

$$\begin{aligned}
\Delta \sin(ax+b) &= \sin[a(x+1)+b] - sin[ax+b] \\
&= 2\cos\left(ax+b+\frac{a}{2}\right)\sin\frac{a}{2} \\
&= 2\sin\left(\frac{\pi}{2}+ax+b+\frac{a}{2}\right)\sin\frac{a}{2} \\
&= 2\sin\frac{a}{2}\sin\left((ax+b)+\frac{\pi}{2}+\frac{a}{2}\right) \\
\Delta^2 \sin(ax+b) &= \Delta\left[2\sin\frac{a}{2}\sin\left((ax+b)+\frac{\pi}{2}+\frac{a}{2}\right)\right] \\
&= 2\sin\frac{a}{2}\left[\sin\left\{(a(x+1)+b)+\frac{\pi}{2}+\frac{a}{2}\right\} - \sin\left\{(ax+b)+\frac{\pi}{2}+\frac{a}{2}\right\}\right] \\
&= 2\sin\frac{a}{2}2\cos\left(ax+b+a+\frac{\pi}{2}\right)\sin\frac{a}{2} \\
&= \left(2\sin\frac{a}{2}\right)^2 \sin\left(ax+b+2.\frac{a+\pi}{2}\right)
\end{aligned}$$

Proceeding next steps, we get

$$\Delta^n \sin(ax+b) = \left(2\sin\frac{a}{2}\right)^n \sin\left(ax+b+n.\frac{a+\pi}{2}\right)$$

□

Theorem 14. *If the values of independent variable are equidistant, then the nth differences of polynomial of nth degree are constant and all the higher order differences are zero.*

Proof. Let

$$f(x) = ax^n + bx^{n-1} + \ldots + px + q$$

Then first forward difference is

$$\begin{aligned}
\Delta f(x) &= f(x+h) - f(x) \\
&= [a(x+h)^n + b(x+h)^{n-1} + \ldots + p(x+h) + q] - [ax^n + bx^{n-1} + \ldots + px + q] \\
&= a[(x+h)^n - x^n] + b[(x+h)^{n-1} - (x)^{n-1}] + \ldots + p[x+h-x] \\
&= anhx^{n-1} + b'x^{n-2} + \ldots + p'x + l'
\end{aligned}$$

where $b', c', \ldots$ etc. are new constants.

From above result, it is obvious that the first difference of polynomial of nth degree is a polynomial again and with degree $(n-1)$. Similarly Second difference would be polynomial of $(n-2)$th degree. Continuing this process for nth differences, we get a polynomial of $(n-n) = 0$ degree. i.e. constant polynomial. i.e.

$$\Delta^n f(x) = a.n!.h^n$$

Since nth difference is constant, the $(n+1)$th and higher difference of a polynomial will be zero. Converse of this theorem is also true. □

■ EXAMPLE 17.4

Evaluate $\Delta^n \left[\frac{1}{x}\right]$, given $h = 1$.

Solution: By definition of forward difference we have

$$\Delta f(x) = f(x+1) - f(x)$$

where, we had taken $h = 1$. This implies that

$$\begin{aligned}
&\Delta\left[\frac{1}{x}\right] = \frac{1}{x+1} - \frac{1}{x} = -\frac{1}{x(x+1)} \\
\Rightarrow &\Delta^2\left[\frac{1}{x}\right] = -\Delta\left[\frac{1}{x(x+1)}\right] = -\frac{1}{(x+1)(x+2)} - \frac{1}{x(x+1)} = (-1)^2\frac{1}{x(x+1)(x+2)}
\end{aligned}$$

Proceeding $n-2$ more times as above, we get

$$\Delta\left[\frac{1}{x}\right] = (-1)^n \frac{1}{x(x+1)(x+2)\ldots(x+n)}$$

□

Problems

17.1 Construct the forward difference table, given that

x:	5	10	15	20	25	30
y:	9962	9848	9659	9397	9063	8660

and point out the values of $\Delta^2 y_{10}$, $\Delta^4 y_5$.

17.2 Construct a backward difference table for $y = \log x$ given that

x:	10	20	30	40	50
y:	1	1.3010	1.4771	1.6021	1.6990

and find values of $\nabla^3 \log 40$ and $\nabla^4 \log 50$

17.3 Evaluate (h is the interval of differencing where not given)

a) $\frac{\Delta^2 x^2}{E(x+\log x)}$; $h = 1$

b) $\Delta\left(\frac{e^x}{e^x+e^{-x}}\right)$

c) $\Delta \tan^{-1} x$

d) $\Delta^2\left(\dfrac{5x+12}{x^2+5x+6}\right)$; the interval of differencing being unity.

e) $\Delta(e^{ax} logbx)$

17.4 Assuming that the following values of y belong to a polynomial of degree 4, compute the next three values:

x:	0	1	2	3	4	5	6	7
y:	1	-1	1	-1	1	-	-	-

17.5 Prove that

a) $\Delta \log f(x) = \log\left[1+\frac{\Delta f(x)}{f(x)}\right]$

b) $e^x = \left(\dfrac{\Delta^2}{E}\right)e^x \cdot \frac{Ee^x}{\Delta^2 e^x}$

c) $hD = -\log(1-\nabla) = \sinh^{-1}(\mu\delta)$

d) $(E^{\frac{1}{2}} + E^{\frac{-1}{2}})(1+\Delta)^{\frac{1}{2}} = 2+\Delta$

e) $\Delta = \dfrac{1}{2}\delta^2 + \delta\sqrt{1+\dfrac{\delta^2}{4}}$

f) $\Delta + \nabla = \dfrac{\Delta}{\nabla} - \dfrac{\nabla}{\Delta}$

g) $\Delta^r y_k = \nabla^r y_{k+r}$

h) $(1+\Delta)(1-\nabla) = 1$

i) $\delta[f(x)g(x)] = \mu f(x)\delta g(x) + \mu g(x)\delta f(x)$

j) $\delta\left[\dfrac{f(x)}{g(x)}\right] = \dfrac{\mu g(x)\delta f(x) - \mu f(x)\delta g(x)}{g(x-\frac{1}{2})g(x+\frac{1}{2}}$

k) $\mu\left[\dfrac{f(x)}{g(x)}\right] = \dfrac{\mu f(x)\mu g(x) - \frac{1}{4}\delta f(x)\delta g(x)}{g(x)-\frac{1}{2}g(x)+\frac{1}{2}}$

l) $\mu\delta = \frac{1}{2}(\Delta+\nabla)$

m) $\nabla - \Delta = -\nabla\Delta$

n) $u_4 = u_0 + 4\Delta u_0 + 6\Delta^2 u_{-1} + 10\Delta^3 u_{-1}$

o) $\Delta \sin^{-1} x = \sin^{-1}[(x+1)\sqrt{1-x^2} - x\sqrt{1-(x+1)^2}]$

17.6 Obtain the first term of the series whose second and subsequent terms are $8, 3, 0, -1, 0$

17.7 Given u_0, u_1, u_2, u_3, u_4 and u_5 and assuming the fifth order differences to be constant, prove that

$$u_{2\frac{1}{2}} = \frac{c}{2} + \frac{25(c-b)+3(a-c)}{256}$$

where $a = u_0 + u_5$, $b = u_1 + u_4$, $c = u_2 + u_3$.

17.8 Find the function whose first difference is e^x

17.9 If $\Delta^3 u_x = 0$ prove that

$$u_{x+\frac{1}{2}} = \frac{1}{2}(u_x + u_{x+1}) - \frac{1}{16}(\Delta^2 u_x + \Delta^2 u_{x+1})$$

17.10 Find $\Delta^{10}(1-ax)(1-bx^2)(1-cx^3)(1-dx^4)$

17.11 u_x is a function of x for which fifth differences are constant and $u_1 + u_7 = -786$, $u_2 + u_6 = 686$, $u_3 + u_5 = 1088$ Find u_4.

17.4 FACTORIAL NOTATION

The product

$$x(x-h)(x-2h)...(x-(r-1)h)$$

is known as factorial polynomial and is denoted by $[x]^r$ or $x^{(r)}$

17.4.1 Differences of $x^{(r)}$

$$\begin{aligned}
\Delta x^{(r)} &= (x+h)^{(r)} - x^{(r)} \\
&= (x+h)x(x-h)(x-2h)...(x-(r-2)h) \\
&\quad - x(x-h)(x-2h)...(x-(r-1)h) \\
&= x(x-h)(x-2h)...(x-(r-2)h)[(x+h)-(x-(r-1)h)] \\
&= x(x-h)(x-2h)...(x-(r-2)h)[rh] \\
&= rhx^{(r-1)}
\end{aligned}$$

Similarly,

$$\Delta^2 x^{(r)} = r(r-1)h^2 x^{(r-2)}$$

Proceeding on, we get

$$\Delta^r x^{(r)} = r!h^r$$

This implies

$$\Delta^{r+1} x^{(r)} = 0$$

Remark: The result of differencing of $x^{(r)}$ is analogous to differentiating x^r.

■ **EXAMPLE 17.5**

Express $f(x) = x^4 - 12x^3 + 24x^2 - 30x + 9$ and its successive differences in factorial notation. Hence show that $\Delta^5 f(x) = 0$.

Solution: Let factorial notation for $f(x)$ is

$$f(x) = Ax^{(4)} + Bx^{(3)} + Cx^{(2)} + Dx^{(1)} + E$$

Using method of synthetic divison,we divide by $x,(x-1),(x-2),(x-3)$ etc. successively, then

	1	-12	24	-30	**5=E**
1	0	1	-11	13	
	1	-11	13	**-17=D**	
2	0	2	-18		
	1	-9	**-5=C**		
3	0	3			
	1	**-6=B**			
4	0				
	1=A				

Hence

$$\begin{aligned}
f(x) &= x^{(4)} - 6x^{(3)} - 5x^{(2)} - 17x^{(1)} + 9 \\
\Delta f(x) &= 4x^{(3)} - 18x^{(2)} - 10x^{(1)} - 17 \\
\Delta^2 f(x) &= 12x^{(2)} - 36x^{(1)} - 10 \\
\Delta^3 f(x) &= 24x^{(1)} - 36 \\
\Delta^4 f(x) &= 24 \\
\Delta^5 f(x) &= 0
\end{aligned}$$

□

■ EXAMPLE 17.6

Obatain the function whose first difference is $9x^2 + 11x + 5$.

Solution: Let $f(x)$ be the required function, so that

$$\Delta f(x) = 9x^2 + 11x + 5$$

Let factorial notation of $\Delta f(x)$ be as[1]

$$\Delta f(x) = Ax^{(2)} + Bx^{(1)} + C \qquad = Ax(x-1) + Bx + C = Ax^2 - (A-B)x + C$$

That is

$$9x^2 + 11x + 5 = Ax^2 - (A-B)x + C$$

On comparing, we get

$$A = 9, \;\; B = 20, \;\; C = 5$$

Thus

$$\Delta f(x) = 9x^{(2)} + 20x^{(1)} + 5$$

[1]The vaue of A, B, C can also be obtained as previous example.

Now, we can simply integrate this equation wrt x, as we know that difference operation on factorial notation is equivalent as differential operation:

$$\begin{aligned} f(x) &= 9\frac{x^{(3)}}{3} + 20\frac{x^{(2)}}{2} + 5x^{(1)} + K \\ &= 3x(x-1)(x-2) + 10x(x-1) + 5x + K \\ &= 3x^3 + x^2 + x + K \end{aligned}$$

where K is the constant of integration. □

Problems

17.12 Express $y = 2x^3 - 3x^2 + 3x - 10$ in factorial notation and hence show that $\Delta^3 y = 12$

17.13 Express $f(x) = x^4 - 12x^3 + 24x^2 - 30x + 9$ and its successive differences in factorial notation. Hence show that $\Delta^5 f(x) = 0$

17.14 Obtain the function whose first difference is $9x^2 + 11x + 5$

17.15 Express $f(x) = \frac{x-1}{(x+1)(x+3)}$ in terms of negative factorial polynomials.

17.16 Find the relation between α, β and γ in order that $\alpha + \beta x + \gamma x^2$ may be expressible in one term in the factorial notation.

17.17 Represent the following polynomials
a) $11x^4 + 5x^3 + x - 15$
b) $2x^3 - 3x^2 + 3x + 10$
and its successive differences in factorial notation.

17.5 PROBLEMS OF MISSING TERMS

■ EXAMPLE 17.7

Estimate the production for 1964 and 1966 from the following data:

Year	1961	1962	1963	1964	1965	1966	1967
Production	200	220	260	-	350	-	430

Solution: Since five figures are known, assume all the fifth order differences as zero. Since two figures are unknown, we need two equations to determine them -
Hence

$$\Delta^5 y_0 = 0 \quad \text{and} \quad \Delta^5 y_1 = 0$$
$$\Rightarrow (E-1)^5 y_0 = 0 \quad \text{and} \quad (E-1)^5 y_1 = 0$$

Thse equation gives, respectively,

$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0 y_6 - 5y_5 + 10y_4 - 10y_3 + 5y_2 - y_1 \qquad = 0$$

Substituting the known values, we get

$$y_5 - 1750 + 10y_3 - 2600 + 1100 - 200 = 0$$
$$430 - 5y_5 + 3500 - 10y_3 + 1300 - 220 = 0$$

This implies

$$y_5 + 10y_3 = 3450$$
$$-5y_5 - 10y_3 = -5010$$

On solving these equations, we get

$$y_3 = 306 \quad \text{and} \quad y_5 = 390$$

Hence the production of the year 1964 is 364 and production of the year 1966 is 390. □

Problems

17.18 Given, $\log 100$, $\log 101 = 2.0043$, $\log 103 = 2.0128$, $\log 104 = 2.0170$ Find $\log 102$.

17.19 A second degree polynomial passes through $(0,1), (1,3), (2,7), (3,13)$. Find the polynomial.

17.20 Estimate the production for 1964 and 1966 from the following data :

year	1961	1962	1963	1964	1965	1966	1967
Production	200	220	260	-	350	-	430

17.21 Find the missing value of the following data :

x	1	2	3	4	5
$f(x)$	7	x	13	21	37

17.22 Find the missing terms in the following table :

x	1	2	3	4	5	6	7	8
$f(x)$	1	8	?	64	?	216	343	512

17.6 NEWTON'S FORWARD INTERPOLATION

Let the function $y = f(x)$ take the values $y_0, y_1, y_2, \ldots, y_n$ corresponding to the values $x_0, x_1, x_2, \ldots, x_n$. Where $x_i = x_0 + ih$, $i = 0, 1, 2, \ldots, n$. Suppose it is required to evaluate $f(x)$ for the $x = x_0 + ph$, where p is any real number.

We have for any real number p, we have defined E such that

$$E^p f(x) = f(x_0 + ph)$$

That is

$$y_p = E^p y_0 = (1+\Delta)^p y_0$$
$$= \left[1 + p\Delta + \frac{p(p-1)}{2!}\Delta^2 + \frac{p(p-1)(p-2)}{3!}\Delta^2 + ...\right] y_0$$

(Using Binomial Theorem)

Thatis,

$$y_p = y_0 + p\Delta y_0 + \frac{p(p-1)}{2!}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{3!}\Delta^3 y_0 + ...$$

If $y = f(x)$ is a polynimial of nth degree, then Δ^{n+1} and higher differences will be zero. Hence

$$y_p = y_0 + p\Delta y_0 + \frac{p(p-1)}{2!}\Delta^2 y_0 + ... + \frac{p(p-1)...(p-n+1)}{n!}\Delta^n y_0 \qquad (17.5)$$

Newton's forward formula .

EXAMPLE 17.8

Using Newton's forward interpolation formula,find the area of a circle of diameter 82 metre from the given table of diameter and area of circle:

Diameter(metre)	80	85	90	95	100
Area (metre^2)	5026	5674	6362	7088	7854

Solution: The forward difference table is as follows:

Diameter(x)	Area(y)	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
80	5026				
		648			
85	5674		40		
		688		-2	
90	6362		38		4
		726		2	
95	7088		40		
		766			
100	7854				

$$p = \frac{x - x_0}{h} = \frac{82-80}{5} = \frac{2}{5} = 0.4$$

Applying Newton's forward difference formula, we have

$$\begin{aligned}
y &= y_0 + p\Delta y_0 + \frac{p(p-1)}{2!}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{3!}\Delta^3 y_0 \\
&+ \frac{p(p-1)(p-2)(p-3)}{4!}\Delta^4 y_0 \\
&= 5026 + 0.4\times 648 + \frac{(0.4)(0.4-1)}{2}\times 40 + \frac{0.4(0.4-1)(0.4-2)}{6}\times -2 \\
&+ \frac{0.4(0.4-1)(0.4-2)(0.4-3)}{24}\times 4 \\
&= 5026 + 259.2 - 4.8 - 0.128 - 0.416 \\
&= 5279.856
\end{aligned}$$

□

■ EXAMPLE 17.9

Find the cubic polynomial which takes the following values:

x:	0	1	2	3
$f(x)$:	1	2	1	10

Solution: Let us form the difference table:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
0	**1**			
		1		
1	2		**-2**	
		-1		**12**
2	1		10	
		9		
3	10			

Here, $h = 1$. Hence using the formula, $x = x_0 + ph$, and choosing $x_0 = 0$,we get

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

∴ By Newton's forward difference formula,

$$\begin{aligned}
y &= y_0 + x\Delta y_0 + \frac{x(x-1)}{2!}\Delta^2 y_0 + \frac{x(x-1)(x-2)}{3!}\Delta^3 y_0 \\
&= 1 + x(1) + \frac{x(x-1)}{2!}(-2) + \frac{x(x-1)(x-2)}{3!}(12) \\
&= 2x^3 - 7x^2 + 6x + 1
\end{aligned}$$

Hence the required cubic polynomial is $y = 2x^3 - 7x^2 + 6x + 1$. □

■ EXAMPLE 17.10

The following table gives the marks secured by 100 students in the Numerical Analysis subject:

Range of Marks:	30-40	40-50	50-60	60-70	70-80
No. of students:	25	35	22	11	7

Use Newton's forward difference interpolation formula to find.
(i) the number of students who got more than 55 marks.
(ii)the number of students who secured marks in the range from 36 to 45.

Solution: The given table is re-arranged as follows:

Marks obtained	No.of students
Less than 40	25
Less than 50	60
Less than 60	82
Less than 70	93
Less than 80	100

The difference table is as under:

Marks obtained less than	No.of students=y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
40	**25**				
		35			
50	60		**-13**		
		22		**2**	
60	82		-11		**5**
		11		7	
70	93		-4		
		7			
80	100				

(i) Here, $x_0 = 40$, $h = 10$, $x_0 + ph = 55$,Therefore $40 + 10p = 55$

$$\rightarrow p = 1.5$$

First, we find the number of students who got less than 55 marks.
Applying Newton's forward difference formula,

$$\begin{aligned} y_{55} &= y_{40} + p\Delta y y_{40} + \frac{p(p-1)}{2!}\Delta^2 y_{40} + \frac{p(p-1)(p-2)}{3!}\Delta^3 y_{40} \\ &+ \frac{p(p-1)(p-2)(p-3)}{4!}\Delta^4 y_{40} \\ &= 25 + (1.5)(35) + \frac{(1.5)(0.5)}{2!}(-13) \\ &+ \frac{(1.5)(.5)(-.5)}{3!}(2) \\ &+ \frac{(1.5)(.5)(-.5)(-1.5)}{4!}(5) \\ &= 71.6171875 = 72 \end{aligned}$$

There are 72 students who got less than 55 marks. i.e., No. of students who got more than 55 marks = 100-72 =28

(ii) To calculate the number of students securing marks between 36 and 45,take the difference of y_{45} and y_{36}. Here again

$$p_{36} = \frac{x - x_0}{h} = \frac{36 - 40}{10} = -0.4$$

Also,

$$p = \frac{45 - 40}{10} = 0.5$$

By Newton's forward difference formula.

$$\begin{aligned}
y_{36} &= y_{40} + p\Delta y_{40} + \frac{p(p-1)}{2!}\Delta^2 y_{40} + \frac{p(p-1)(p-2)}{3!}\Delta^3 y_{40} \\
&+ \frac{p(p-1)(p-2)(p-3)}{4!}\Delta^4 y_{40} \\
&= 25 + (-.4)(35) + \frac{(-.4)(-1.4)}{2!}(-13) + \frac{(-.4)(-1.4)(-2.4)}{3!}(2) \\
&+ \frac{(-.4)(-1.4)(-2.4)(-3.4)}{4!}(5) = 7.864 \approx 8
\end{aligned}$$

$$\begin{aligned}
\text{Also}, y_{45} &= y_{40} + p y_{40} + \frac{p(p-1)}{2!}\Delta^2 y_{40} \\
&+ \frac{p(p-1)(p-2)}{3!}\Delta^3 y_{40} + \frac{p(p-1)(p-2)(p-3)}{4!}\Delta^4 y_{40} \\
&= 25 + (.5)(-.5) + \frac{(.5)(-.5)}{2}(-13) + \frac{(.5)(-.5)(-1.5)}{6}(2) + \frac{(.5)(-.5)(-1.5)(-2.5)}{24}(5) \\
&= 44.0546 \approx 44
\end{aligned}$$

Hence the number of students who secured marks in the range from 36 to 45 is

$$y_{45} - y_{36} = 44 - 8 = 36$$

□

Problems

17.23 Find the value of $\sin 52^o$ from the given table :

θ^o	45^o	50^o	55^o	60^o
$\sin\theta$	0.7071	0.7660	0.8192	0.8660

17.24 From the following table, find the value of $e^{0.24}$

x	0.1	0.2	0.3	0.4	0.5
e^x	1.10517	1.22140	1.34986	1.49182	1.64872

17.25 From the following table of half-yearly premium for policies maturing at different ages, estimate the premium for policies maturing at age of 46

Age	45	50	55	60	65
Premium (in Rs.)	114.84	96.16	83.32	74.48	68.48

17.26 From the table, estimate the number of students who obtained marks between 40 and 45

Marks	30-40	40-50	50-60	60-70	70-80
No. of students	31	42	51	35	31

17.27 Find the cubic polynomial which takes the following values :

x	0	1	2	3
$f(x)$	1	2	1	10

17.28 The following are the numbers of deaths in four successive ten year age groups. Find the number of deaths at $45-50$ and $50-55$.

Age group	25-35	35-45	45-55	55-65
Deaths	13229	18139	24225	31496

17.29 If p, q, r, s be the successive entries corressponding to equidistant arguments in a table, show that when third differences are taken into account, the entry corresponding to the argument half way between the arguments at q and r is $A + \left(\frac{B}{24}\right)$, where A is the arithmetic mean of q and r and B is arithmetic mean of $3q-2p-s$ and $3r-2s-p$.

17.30 The following table gives the distance in nautical miles of the visible horizon for the given heights in feet above the earth's surface.

x	100	150	200	250	300	350	400
y	10.63	13.03	15.04	16.81	18.42	19.9	21.27

Use Newton's forward formula to find y when $x = 218$ ft.

17.31 If l_x represents the number of persons living at age x in a life table, find as accurately as the data will permit l_x for values of $x = 35, 42$ and 47. Given l_{20}=512, $l_{30} = 390, l_{40} = 360, l_{50} = 243$

17.32 Given that:

x	1	2	3	4	5	6
$y(x)$	0	1	8	27	64	125

Find the value of $f(2.5)$

17.33 Use Newton's forward difference formula to obtain the interpolating polynomial $f(x)$ satisfying the following data :

x	1	2	3	4
$f(x)$	26	18	4	1

17.34 In the following table, values of y are consecutive terms of a series of which 23.6 is the 6th term. Find the first and tenth terms of the series.

x	3	4	5	6	7	8	9
y	4.8	8.4	14.5	23.6	36.2	52.8	73.9

17.7 NEWTON'S BACKWARD INTERPOLATION

Let the function $y = f(x)$ take the values $y_0, y_1, y_2, \ldots, y_n$ corresponding to the values $x_0, x_1, x_2, \ldots, x_n$. Where $x_i = x_0 + ih$, $i = 0, 1, 2, \ldots, n$. Suppose it is required to evaluate $f(x)$ for the $x = x_0 + ph$, where p is any real number.

We have for any real number p, we have defined E such that

$$E^p f(x) = f(x_n + ph)$$

That is

$$y_p = E^p y_n = \left[\frac{1}{1-\nabla}\right]^p y_n = (1-\nabla)^{-p} y_n$$
$$= \left[1 + p\nabla + \frac{p(p+1)}{2!}\nabla^2 + \frac{p(p+1)(p+2)}{3!}\nabla^3 + \ldots\right] y_n$$

(Using Binomial Theorem)

$Thatis,$

$$y_p = y_n + p\nabla y_n + \frac{p(p+1)}{2!}\nabla^2 y_n + \frac{p(p+1)(p+2)}{3!} y_n + \ldots$$

If $y = f(x)$ is a polynimial of nth degree, then ∇^{n+1} and higher differences will be zero. Hence

$$y_p = y_n + p\nabla y_n + \frac{p(p+1)}{2!}\nabla^2 y_n + \ldots + \frac{p(p+1)\ldots(p+n-1)}{n!}\nabla^n y_n \qquad (17.6)$$

This formula is known as Newton's backward formula..

■ EXAMPLE 17.11

The table gives the distance in nautical miles of the visible horizon for the given height in feet above the earth's surface:

x=height :	200	250	300	350	400
y=distance :	15.04	16.81	18.42	19.90	21.27

Find the values of y when x=410 ft.

Solution: The difference table is

x	y	Δ	Δ^2	Δ^3	Δ^4
200	15.04				
		1.77			
250	16.81		-0.16		
		1.61		0.03	
300	18.42		-0.13		**-0.01**
		1.48		**0.02**	
350	19.90		**-0.11**		
		1.37			
400	**21.27**				

Since $x = 400$ is near the end of the table, we use Newton's backward interpolation formula.

$$y_p = y_n + p\nabla y_n + \frac{p(p+1)}{2!}\nabla^2 y_n + \ldots + \frac{p(p+1)\ldots(p+n-1)}{n!}\nabla^n y_n$$

Here $x_n = 400$, and

$$p = \frac{x - x_n}{h} = \frac{410 - 400}{} = \frac{10}{50} = 0.2$$

Thus required value

$$\begin{aligned}
y_{410} &= y_4 00 + p\nabla y_{400} + \frac{p(p+1)}{2!}\nabla^2 y_{400} \\
&+ \frac{p(p+1)(p+2)}{3!}\nabla^3 y_{400} + \frac{p(p+1)(p+2)(p+3)}{4!}\nabla^4 y_{400} \\
&= 21.27 + (0.2)(1.37) + \frac{(0.2)(1.2)}{2!}(-0.11) \\
&+ \frac{(0.2)(1.2)(2.2)}{3!}(0.02) + \frac{(0.2)(1.2)(2.2)(3.2)}{4!}(-0.01) \\
&= 21.27 + 0.274 + 0.0132 + 0.0018 + -0.0007 \qquad = 21.53 \text{ nautical miles.}
\end{aligned}$$

□

Problems

17.35 Given the following table:

x	40	45	50	55	60	65
$\log x$	1.60206	1.65321	1.69897	1.74036	1.77815	1.81291

Find the value of $\log 58.75$.

17.36 The probability intergral $P = \sqrt{\frac{2}{\pi}}\int_0^x e^{\frac{-}{1}} 2t^2 dt$ has following values:

x	1.00	1.05	1.10	1.15	1.20	1.25
P	0.682689	0.706282	0.728668	0.749856	0.769861	0.788700

Calculate P for $x = 1.235$.

17.37 In an examination, the number of candidates who obtained marks between certain limits are as follows:

Marks	0-19	20-39	40-59	60-79	80-99
No. of candidates	41	62	65	50	17

Estimate the number of candidates who obtained fewer than 70 marks.

17.38 Using Newton's backward difference formula, find the value of $e^{-1.9}$ from the following table of values of e^{-x}:

x	1	1.25	1.50	1.75	2.00
e^{-x}	0.3679	0.2865	0.2231	0.1738	0.1353

17.39 Calculate the value of $\tan 48^o 15'$ from the following table:

x^o	45	46	47	48	49	50
$\tan x^o$	1.00000	1.03053	1.07237	1.11061	1.15037	1.19175

17.40 From the following table of half yearly premium for policies maturing at different ages, estimate the premium for policy maturing at the age of 63:

Age	45	50	55	60	65
Premium (in Rs.)	114.84	96.16	83.32	74.48	68.48

17.8 INTERPOLATION: UNEQUALLY SPACED POINTS

In this section, we consider the cases where data corressponding to independent variable is unequally spaced. To interpolate such function, here, we discuss two formulae:

1. Lagrange's Interpolation Formula
2. Newton's Divided Difference Interpolation Formula

17.8.1 Lagranges Interpolation

Let $f(x_0), f(x_1), ..., f(x_n)$ be $n+1$ entries of a function $y = f(x)$, where $f(x)$ is assumed too be polynomial corresponding to the arguments $x_0, x_1, ..., x_n$.
The polynomial $f(x)$ may be written as

$$\begin{aligned} f(x) = A_0(x-x_1)(x-x_2)...(x-x_n) \\ +A_1(x-x_0)(x-x_2)...(x-x_n)+... \\ +A_n(x-x_0)(x-x_1)...(x-x_{n-1}) \end{aligned}$$

where $A_0, A_1, ..., A_n$ are constants to be determined.

Putting $x = x_0$

$$f(x_0) = A_0(x_0-x_1)(x_0-x_2)...(x_0-x_n)$$

$$A_0 = \frac{f(x_0)}{(x_0-x_1)(x_0-x_2)...(x_0-x_n)}$$

Putting $x = x_1$

$$f(x_1) = A_1(x_1-x_0)(x_1-x_2)...(x_1-x_n)$$

$$A_1 = \frac{f(x_1)}{(x_1-x_1)(x_1-x_2)...(x_1-x_n)}$$

Similarly,

$$A_n = \frac{f(x_n)}{(x_n-x_0)(x_n-x_1)...(x_n-x_{n-1})}$$

Hence, on substituting the constants values, we get

$$\begin{aligned} f(x) = \frac{(x-x_1)(x-x_2)...(x-x_n)}{(x_0-x_1)(x_0-x_2)...(x_0-x_n)} f(x_0) \\ + \frac{(x-x_0)(x-x_2)...(x-x_n)}{(x_1-x_1)(x_1-x_2)...(x_1-x_n)} f(x_1) + ... \\ + \frac{(x-x_0)(x-x_1)...(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)...(x_n-x_{n-1})} f(x_n) \end{aligned}$$

Or

$$f(x) = \sum_{r=0}^{n} \frac{(x-x_0)(x-x_1)...(x-x_{r-1})(x-x_{r+1})...(x-x_{n-1})}{(x_r-x_0)(x_r-x_1)...(x_r-x_{r-1})(x_r-x_{r+1})...(x_r-x_{n-1})} f(x_r)$$

This result is known as Lagrange's Interpolation Formula.

■ **EXAMPLE 17.12**

Prove that Lagrange's formula can be put in the for $f(x) = \sum_{r=0}^{n} \frac{\phi(x) f(x_r)}{(x-x_r)\phi'(x_r)}$, where $\phi(x) = \prod_{r=0}^{n}(x-x_r)$ and $\phi'(x_r) = \left[\frac{d\phi(x)}{dx}\right]_{x=x_r}$

Solution: We have Lagrange's formula,

$$f(x)=\sum_{r=0}^{n}\frac{(x-x_0)(x-x_1)\ldots(x-x_{r-1})(x-x_{r+1})\ldots(x-x_{n-1})}{(x_r-x_0)(x_r-x_1)\ldots(x_r-x_{r-1})(x_r-x_{r+1})\ldots(x_r-x_{n-1})}f(x_r)$$
$$=\sum_{r=0}^{n}\left[\frac{\phi(x)}{x-x_r}\right]\left[\frac{f(x_r)}{(x_r-x_0)(x_r-x_1)\ldots(x_r-x_{r-1})(x_r-x_{r+1})\ldots(x_r-x_{n-1})}\right]$$

Now,

$$\begin{aligned}
\phi(x)=\phi(x)&=\prod_{r=0}^{n}(x-x_r)\\
&=(x-x_0)(x-x_1)\ldots(x-x_{r-1})(x-x_r)(x-x_{r+1})\ldots(x-x_{n-1})\\
\Rightarrow \phi'(x)&=(x-x_1)(x-x_2)\ldots(x-x_r)\ldots(x-x_n)\\
&\quad+(x-x_0)(x-x_2)\ldots(x-x_r)\ldots(x-x_n)+\ldots\\
&\quad\quad+(x-x_0)(x-x_1)\ldots(x-x_r)\ldots(x-x_{n-1})\\
\Rightarrow \phi'(x)&=[(x-x_1)(x-x_2)\ldots(x-x_r)\ldots(x-x_n)\\
&\quad+(x-x_0)(x-x_2)\ldots(x-x_r)\ldots(x-x_n)+\ldots\\
&\quad\quad+(x-x_0)(x-x_1)\ldots(x-x_r)\ldots(x-x_{n-1})]_{x=x_r}\\
&=(x_r-x_0)(x_r-x_1)\ldots(x_r-x_{r-1})(x_r-x_{r+1})\ldots(x_r-x_{n-1})
\end{aligned}$$

Hence,

$$f(x)=\sum_{r=0}^{n}\frac{\phi(x)f(x_r)}{(x-x_r)\phi'(x_r)}$$

□

■ EXAMPLE 17.13

The function $y=\sin x$ is tabulated below: Using the Lagrange's interpolation formula, find the

x	0	$\pi/4$	$\pi/2$
$\sin x$	0	0.70711	1.0

value of $\sin(\pi/6)$.

Solution: We have,

$$\begin{aligned}
\sin\frac{\pi}{6}&=\frac{\left(\frac{\pi}{6}-0\right)\left(\frac{\pi}{6}-\frac{\pi}{2}\right)}{\left(\frac{\pi}{4}-0\right)\left(\frac{\pi}{4}-\frac{\pi}{2}\right)}(0.70711)+\frac{\left(\frac{\pi}{6}-0\right)\left(\frac{\pi}{6}-\frac{\pi}{4}\right)}{\left(\frac{\pi}{2}-0\right)\left(\frac{\pi}{2}-\frac{\pi}{4}\right)}(1)\\
&=0.51743
\end{aligned}$$

□

17.8.2 Newton's Divided Difference Interpolation Formula

The Lagrange's formula has the drawback that if another interpolation values were inserted then the interpolation coefficients are required to be recalculated. This labour of recomputing is saved by using Newton's general interpolation formula which is based on divided differences.

17.8.3 Divided difference

If (x_0,y_0), (x_1,y_1), (x_2,y_2), ... be given points, then the first divided difference for the argument x_0,x_1 is denoted as $[x_0,x_1]$ defined as

$$[x_0,x_1] = \frac{y_1-y_0}{x_1-x_0}$$

Similarly, $[x_1,x_2] = \frac{y_2-y_1}{x_2-x_1}$ and $[x_2,x_3] = \frac{y_3-y_2}{x_3-x_2}$ etc.
The second divided difference for x_0,x_1,x_2 is defined as

$$[x_0,x_1,x_2] = \frac{[x_1,x_2]-[x_0,x_1]}{x_2-x_0}$$

The third divided difference for x_0,x_1,x_2,x_3 is defined as

$$[x_0,x_1,x_2,x_3] = \frac{[x_1,x_2,x_3]-[x_0,x_1,x_2]}{x_3-x_0}$$

and so on.

Here it is notable that we can write $[x_0,x_1]$ as

$$[x_0,x_1] = \frac{y_0}{x_0-x_1} + \frac{y_1}{x_1-x_0} = [x_1,x_0]$$

which shows symmetry in divided differences. Thus, the divided are symmetrical in their arguments, i.e. independent of the order of the arguments.

■ EXAMPLE 17.14

Show that, the *n*th divided difference of a polynomial of the *n*th degree is constant.

Solution: Let the arguments be equally spaced so that $x_1-x_0 = x_2-x_1 = \ldots = h$. Then

$$[x_0,x_1] = \frac{y_1-y_0}{x_1-x_0} = \frac{\triangle y_0}{h}$$

$$[x_0,x_1,x_2] = \frac{[x_1,x_2]-[x_0,x_1]}{x_2-x_0} = \frac{1}{2h}\left\{\frac{\triangle y_1}{h} - \frac{\triangle y_1}{h}\right\} = \frac{1}{2!h^2}\Delta^2 y_0$$

and in general,

$$[x_0,x_1,x_2.....,x_n] = \frac{1}{n!h^n}\triangle^n y_0.$$

If the tabulated function is a *n*th degree polynomial, then $\triangle^n y_0$ will be constant. Here the *n*th divided differences will also be constant. □

17.8.4 Newton's General Interpolation Formula

Let $y_0,y_1,.......,y_n$be the value of $y = f(x)$ corresponding to the arguments $x_0,x_1,..........,x_n.$Then from the definition of divided differences,we have

$$[x,x_0] = \frac{y-y_0}{x-x_0}$$

So that,

$$y = y_0 + (x-x_0)[x,x_0] \tag{17.7}$$

Again,

$$[x,x_0,x_1] = \frac{[x,x_0]-[x_0,x_1]}{x-x_1}$$

which gives

$$[x, x_0] = [x_0, x_1] + (x - x_1)[x, x_0, x_1]$$

Substituting this value of $[x, x_0]$ in (17.7),we get

$$y = y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x, x_0 x_1] \tag{17.8}$$

Also

$$[x, x_0, x_1, x_2] = \frac{[x, x_0, x_1] - [x_0, x_1, x_2]}{x - x_2}$$

which gives

$$[x, x_0, x_1] = [x_0, x_1, x_2] + (x - x_2)[x, x_0, x_1, x_2]$$

Substituting this value of $[x, x_0, x_1]$in (17.7),we obtain

$$y = y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] + (x - x_0)(x - x_1)(x - x_2)[x, x_0, x_1, x_2]$$

Proceeding in this manner,we get

$$\begin{aligned} y = f(x) &= y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\ &+ (x - x_0)(x - x_1)(x - x_2)[x_0, x_1, x_2, x_3] + \ldots \\ &+ (x - x_0)(x - x_1)\ldots.(x - x_n)[x, x_0, xx_1, \ldots.x_n] \end{aligned}$$

which is called Newton's general interpolation formula with divided differences.

■ EXAMPLE 17.15

Using the following table find $f(x)$ as a polynomial in x.

x	-1	0	3	6	7
$f(x)$	3	-6	39	822	1611

Solution: The divided difference table is

x	$f(x)$	I DD	II DD	III DD	IV DD
-1	**3**				
		-9			
0	-6		**6**		
		15		**5**	
3	39		41		**1**
		261		13	
6	822		132		
		789			
7	1611				

Therefore by Newton's divided difference formula,

$$\begin{aligned} f(x) &= 3 + (x + 1)(-9) + x(x + 1)(6) + x(x + 1)(x - 3)(5) + x(x + 1)(x - 3)(x - 6) \\ &= x^4 - 3x^3 + 5x^2 - 6 \end{aligned}$$

□

EXAMPLE 17.16

Given the values:

x	5	7	11	13	17
f(x)	150	392	1452	2366	5202

Evaluate $f(9)$, using Lagrange's and Newton's divided difference formulae.

Solution:

By Lagrange's formula. Here

$$x_0 = 5, x_1 = 7, x_2 = 11, x_3 = 13, x_4 = 17$$

and

$$y_0 = 150, y_1 = 392, y_2 = 1452, y_3 = 2366, y_4 = 5202$$

On putting $x = 9$ and substituting the above values in Lagrange's formula, we get

$$\begin{aligned} f(9) = {} & \frac{(9-7)(9-11)(9-13)(9-17)}{(5-7)(5-11)(5-13)(5-17)} \times 150 + \frac{(9-5)(9-11)(9-13)(9-17)}{(7-5)(7-11)(7-13)(7-17)} \times 392 \\ & + \frac{(9-5)(9-7)(9-13)(9-17)}{(11-5)(11-7)(11-13)(11-17)} \times 1452 + \frac{(9-5)(9-7)(9-11)(9-17)}{(13-5)(13-7)(13-11)(13-17)} \times 2366 \\ & + \frac{(9-5)(9-7)(9-11)(9-13)}{(17-5)(17-7)(17-11)(17-13)} \times 5202 = -\frac{50}{2} + \frac{3136}{15} + \frac{3872}{3} - \frac{2366}{3} + \frac{578}{5} = 810. \end{aligned}$$

By Newton's divided differences formula. The divided differences (DD) table is

x	y	Ist DD	2nd DD	3rd DD
5	150			
		121		
7	392		24	
		265		1
11	1452		32	
		457		1
13	2366		42	
		709		
17	5202			

Taking $x = 9$ in the Newton's divided difference formula,we obtain

$$\begin{aligned} f(9) &= 150 + (9-5) \times 121 + (9-5(9-7) \times 24 + (9-5(9-7)(9-11) \times 1 \\ &= 150 + 484 + 192 - 16 = 810 \end{aligned}$$

□

Problems

17.41 Given that : $f(0) = 8, f(1) = 68, f(5) = 123$, construct a divided difference table using Newton's divided difference interpolation formula. Deterine the value of f(2).

17.42 Using divided difference, find the value of $f(8)$, given that $f(6) = 1.556$, $f(7) = 1.908$, $f(12) = 2.158$.

17.43 Find a polynomial satisfied by $(-4,1245),(-1,33),(0,5),(2,9)$ and $(5,1335)$, by the use of Newton's interpolation formula with divided differences.

17.44 Using the following table, find $f(x)$ interpolating polynomial:

x	-1	0	2	3	7	10
$f(x)$	-11	1	1	1	141	561

Also find $f'(6), f''(6), f'''(6)$.

17.45 Find Third divided difference of y using the following table:

x	1	3	5	8
y	10	16	26	44

17.46 If $\log 2 = 0.30103, \log 3 = 0.47712, \log 5 = 0.69897, \log 7 = 0.84510$, find the value of $\log 4.7$ to four places of decimal.

17.47 Find the unique polynomial $p(x)$ of degree 2 such that $P(1)$=1, $P(3)$=27, $P(4)$=64 Use Lagrange method of interpolation.

17.48 The function $y = f(x)$ is given at the points $(7,3),(8,1),(9,1)$ and $(10,9)$. Find the value of y for $x = 9.5$ using Lagrange's interpolation formula.

17.49 By means of Lagrange's formula, prove that

a) $y_0 = \frac{1}{2}(y_1 + y_{-1}) - \frac{1}{8}\left[\frac{1}{2}(y_3 - y_1) - \frac{1}{2}(y_{-1} - y{-3})\right]$

b) $y_3 = 0.05(y_0 + y_6) - 0.3(y_1 + y_5) + 0.75(y_2 + y_4)$

c) $y_1 = y_3 - 0.3(y_5 - y_{-3}) + 0.2(y_{-3} - y_{-5})$

17.50 Four equidistant values u_{-1}, u_0, u_1 and u_2 being given, a value is interpolated by Lagrange's formula, show that it may be written in the form

$$u_x = yu_0 + xu_1 + \frac{y(y^2-1)}{3!}\Delta^2 u_{-1} + \frac{x(x^2-1)}{3!}\Delta^2 u_0$$

where $x + y = 1$.

17.51 Apply Lagrange's formula to find $f(5)$ and $f(6)$ given that $f(2) = 4, f(1) = 2, f(3) = 8, f(7) = 128$. Explain why the result differs from those obtained by completing the series of powers of 2?

17.52 If $y_0, y_1, \ldots . y_9$ are consecutive terms of a series, prove that

$$y_5 = \frac{1}{70}\left[56(y_4 + y_6) - 28(y_3 + y_7) + 8(y_2 + y_8) - (y_1 + y_9)\right]$$

17.53 Applying Lagrange's formula, find a cubic polynomial which approximates the following data :

x	-2	-1	2	3
$y(x)$	-12	-8	3	5

17.54 Find the equation of the cubic curve which passes through the points $(4,-43),(7,83),(9,327)$ and $(12,1053)$

17.55 Values of $f(x)$ are given at a, b and c . Show that the maximum $f(x)$ is obtained at

$$x = \frac{f(a)(b^2 - c^2) + f(b)(c^2 - a^2) + f(c)(a^2 - b^2)}{f(a)(b-c) + f(b)(c-a) + f(c)(a-b)}$$

UNIT 5
NUMERICAL TECHNIQUES - II

CHAPTER 18

SOLVING LINEAR SYSTEM

This chapter covers extremely important problem how to solve a large system of equations? The algebraic system of linear equations occur abundantly in various fields of science and engineering, like elasticity,electrical engineering ,fluid dynamics,heat transfer,structural analysis,statistics, biomedical engineering,vibration engineering and in control theory.A number of problems in numerical analysis can be reduced to, or approximated by, a system of linear equations.

One of the most common source of such systems is through discretization of ordinary or partial differential equations or integral equations. The system of n-linear equations in n-unknowns can be written as:

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_1 + \ldots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_1 + \ldots + a_{2n}x_n &= b_2 \\
&\ldots \\
a_{n1}x_1 + a_{n2}x_1 + \ldots + a_{nn}x_n &= b_n
\end{aligned}
$$

where the a's are constant coefficients,the b's are constants and n is the number of equations.

The above system of linear equations may be expressed in a matrix notation as follows:

$$AX = B$$

where the coefficient matrix A is given by

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}, \ X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \ B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

The augmented matrix is denoted as A_f and is:

$$A_f = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & & a_{2n} & b_2 \\ \vdots & & & & \\ a_{n1} & a_{n2} & & a_{nn} & b_n \end{bmatrix}$$

Here, one may note that if the rank of matrix A is less than the rank of matrix A_f,i.e, $r(A) < r(A_f)$ then the system is said to be inconsistant.

If $r(A) = r(A_f)$,then the system is said to be consistant and has solution. We can find the solution of such equations by direct methods or iterative methods. Among the direct methods, we will discuss Crout's method and in case of iterative methods, we will describe Gauss-seidel.

18.1 CROUT'S METHOD

In this method , the coefficient matrix A is decomposed into matrices L and U. Here the matrix L is lower triangular matrix and U is the unit upper triangular matrix.

Let us consider the system of linear equations:

$$a_{11}x_1 + a_{12}y_2 + a_{13}x_3 = b_1$$
$$a_{21}x_1 + a_{22}y_2 + a_{23}x_3 = b_2$$
$$a_{31}x_1 + a_{32}y_2 + a_{33}x_3 = b_3$$

compare these equation with matrix equation $AX = B$

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \text{ and } B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

Suppose we decomposed $A = LU$ such that

$$L = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \text{ and } U = \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

Since $AX = B$ and $LUX = B$ implies $LY = b$, where $UX = Y$
Such that

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

We have,

$$A = LU$$

That is,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} l_{11} & l_{11}u_{12} & l_{11}u_{13} \\ l_{21} & l_{21}u_{12} + l_{22} & l_{21}u_{13} + l_{22}u_{23} \\ l_{31} & l_{31}u_{12} + l_{32} & l_{31}u_{13} + l_{32}u_{23} + l_{33} \end{bmatrix}$$

Equating coefficients and then on simplification, we get

$$l_{11} = a_{11},\ l_{21} = a_{21},\ l_{31} = a_{31},\ u_{12} = \frac{a_{12}}{a_{11}},\ u_{13} = \frac{a_{13}}{a_{11}}$$

and, then

$$l_{22} = a_{22} - l_{21}u_{12},\ l_{32} = a_{32} - l_{31}u_{12},\ u_{23} = \frac{a_{23} - l_{21}u_{13}}{l_{22}},\ l_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23}$$

Now, we have all the values of L and U. Again, We have

$$LY = B$$

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

which gives

$$\begin{aligned} l_{11} &= b_1 & &\Rightarrow y_1 = \frac{b_1}{l_{11}} \\ l_{21}y_1 + l_{22}y_2 &= b_2 & &\Rightarrow y_2 = \frac{b_2 - l_{21}y_1}{l_{22}} \\ l_{31}y_1 + l_{32}y_2 + l_{33}y_3 &= b_3 & &\Rightarrow y_3 = \frac{b_3 - l_{31}y_1 - l_{32}y_2}{l_{33}} \end{aligned}$$

With the knowledge of Y, L, U and $UX = Y$, we can easily compute X. This procedure is called Crout's Method.

■ EXAMPLE 18.1

Solve the system of linear equations: $x+y+z=1$, $3x+y-3z=5$, $x-2y-5z=10$ by Crout's method.

Solution: Here

$$[A|B] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & : & b_1 \\ a_{21} & a_{22} & a_{23} & : & b_2 \\ a_{31} & a_{32} & a_{33} & : & b_3 \end{bmatrix} \approx \begin{bmatrix} 1 & 1 & 1 & : & 1 \\ 3 & 1 & -3 & : & 5 \\ 1 & -2 & -5 & : & 10 \end{bmatrix}$$

Let the auxillary matrix

$$D = \begin{bmatrix} l_{11} & u_{12} & u_{13} & : & y_1 \\ l_{21} & l_{22} & u_{23} & : & y_2 \\ l_{31} & l_{32} & l_{33} & : & y_3 \end{bmatrix}$$

where

$$l_{11} = a_{11} = 1,\ l_{21} = a_{21} = 3,\ l_{31} = a_{31} = 1$$

$$u_{12} = \frac{a_{12}}{a_{11}} = \frac{1}{1} = 1,\ u_{13} = \frac{a_{13}}{a_{11}} = frac11 = 1,\ y_1 = \frac{b_1}{a_{11}} = \frac{1}{1} = 1$$
$$l_{22} = a_{22} - l_{21}u_{12} = -2,\ l_{32} = a_{32} - l_{31}u_{12} = -3$$
$$u_{23} = \frac{a_{23} - l_{21}u_{13}}{l_{22}} = 3,\ y_2 = \frac{b_2 - l_{21}y_1}{l_{22}} = -1$$
$$l_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23} = 3,\ \ y_3 = \frac{b_3 - y_2 l_{32} - y_1 l_{31}}{l_{33}} = 2$$

Thus, the solution is

$$z = y_3 = 2$$
$$y = y_2 - u_{23}z = -7$$
$$x = y_1 - u_2 z - u_{12}y = 6$$

□

Problems

Solve the following system of linear equations by Crout's method:

18.1 $x + y + z = 3,\ 2x - y + 3z = 16,\ 3x + y - z = -3$

18.2 $2x_1 + 3x_2 + x_3 = -1,\ 5x_1 + x_2 + x_3 = 9,\ 3x_1 + 2x_2 + 4x_3 = 11$

18.3 $2x - y + 3z + w = 9, -x + 2y + z - 2w = 2,$
$3x + y - 4z + 3w = 3, 5x - 4y + 3z - 6w = 2$

18.4 $x_1 + x_2 + x_3 = 1,\ 3x_1 + x_2 - 3x_3 = 5,\ x_1 - 2x_2 - 5x_3 = 10$

18.5 $x + 2y + 3z = 6,\ 2x + 3y + z = 93,\ x + y + 2z = 8$

18.6 $x + y + z = 2, 2x + 3y - 2z = -4, x - 2y + 4z = 17$

18.2 GAUSSSEIDEL METHOD

To understand the GaussSeidel method, we consider a system of 3 linear equations in 3 variables:

$$a_{11}x + a_{12}y + a_{13}z = b_1,\ \ a_{21}x + a_{22}y + a_{23}z = b_2,\ \ a_{31}x + a_{32}y + a_{33}z = b_3$$

These equations may be rewritten as

$$x = \frac{1}{a_{11}}(b_1 - a_{12}y - a_{13}z)$$

$$y = \frac{1}{a_{22}}(b_2 - a_{21}x - a_{23}z)$$
$$z = \frac{1}{a_{33}}(b_3 - a_{31}x - a_{32}y)$$

Now we start with the initial values y_0, z_0 and compute x_1 from the first equation, i.e.,

$$x_1 = \frac{1}{a_{11}}(b_1 - a_{12}y_0 - a_{13}z_0)$$

while using second equation to compute y_1, we use z_0, x_1 for z and x (we are taking x_0)

$$y_1 = \frac{1}{a_{22}}(b_2 - a_{21}x_1 - a_{23}z_0)$$

Now to compute z, for third equation, we use x_1 and y_1, i.e.,

$$z_1 = \frac{1}{a_{33}}(b_3 - a_{31}x_1 - a_{32}y_1)$$

In gereral, this procedure may written as following recurrence relation

$$\begin{aligned} x_{n+1} &= \frac{1}{a_{11}}(b_1 - a_{12}y_n - a_{13}z_n) \\ y_{n+1} &= \frac{1}{a_{22}}(b_2 - a_{21}x_{n+1} - a_{23}z_n) \\ z_{n+1} &= \frac{1}{a_{33}}(b_3 - a_{31}x_{n+1} - a_{32}y_{n+1}) \end{aligned}$$

■ EXAMPLE 18.2

Apply Gauss-Seidal iteration method to solve the equation

$$20x + y - 2z = 17,\ 3x + 20y - z = -18,\ 2x - 3y + 20z = 25$$

Solution: Write the given equation in the form

$$x = \frac{1}{20}(17 - y + 2z) \tag{18.1}$$

$$y = \frac{1}{20}(-18 - 3x + z) \tag{18.2}$$

$$z = \frac{1}{20}(25 - 2x + 3y) \tag{18.3}$$

First iteration,
Putting $y = y_0, z = z_0$ in Equation 18.1, we get

$$x_1 = \frac{1}{20}(17 - y_0 + 2z_0) = 0.8500$$

Putting $x = x_1, z = z_0$ in equation 18.2, we have

$$y_1 = \frac{1}{20}(-18 - 3x_1 + z_0) = -1.0275$$

Putting $x = x_1, y = y_0$ in equation 18.3, we have

$$z_1 = \frac{1}{20}(25 - 2x_1 + 3y_1) = 1.0109$$

Second iteration, Putting $y = y_1, z = z_1$ in 18.1, we get

$$x_2 = \frac{1}{20}(17 - y_1 + 2z_1) = 1.0025$$

Putting$x = x_2, z = z_1$ in 18.2, we obtain

$$y_2 = \frac{1}{20}(-18 - 3x_2 + z_1) = -0.9998$$

Putting $x = x_2, y = y_2$ in 18.3, we get

$$z_2 = \frac{1}{20}(25 - 2x_2 + 3y_2) = 0.9998$$

In third iteration, We get

$$x_3 = \frac{1}{20}(17 - y_2 + 2z_2) = 1.0000$$
$$y_3 = \frac{1}{20}(-18 - 3x_3 + z_2) = -1.0000$$
$$z_3 = \frac{1}{20}(25 - 2x_3 + 3y_3) = 1.0$$

The values in the second and third iterations being practically the same, we can stop. Hence the solution is

$$x = 1,\ y = -1,\ z = 1$$

□

Problems

Solve the following equation by Gauss-Seidel method :

18.7 $x + y + 54z = 110$, $27x + 6y - z = 85$, $6x + 15y + 2z = 72$

18.8 $10x + y + 2z = 44$, $2x + 10y + z = 51$, $x + 2y + 10z = 61$

18.9 $28x + 4y - z = 32$, $x + 3y + 10z = 24$, $2x + 17y + 4z = 35$

18.10 $2x + y + 6z = 9$, $8x + 3y + 2z = 13$, $x + 5y + z = 7$

18.11 $10x + y + z = 12$, $2x + 10y + z = 13$, $2x + 2y + 10z = 14$

18.12 $10x + 2y + z = 9$, $2x + 20y - 2z = -44$, $-2x + 3y + 10z = 22$

18.13 $5x - y + z = 10$, $2x + 4y = 12$, $x + y + 5z = -1$

18.14 $8x - y + z = 18$, $2x + 5y - 2z = 3$, $x + y - 3z = -16$

18.15 $2x + y + z = 4$, $x + 2y - z = 4$, $x + y + 2z = 4$

18.16 $8x + y + z = 8$, $2x + 4y + z = 4$, $x + 3y + 3z = 5$

CHAPTER 19

NUMERICAL DIFFERENTIATION AND INTEGRATION

A given set of $(n+1)$ data points (x_i, y_i), $i = 0,1,2,...,n$ is assumed to represent some function $y = y(x)$. The data can come from some experiment or statistical study, where $y = y(x)$ is unknown, or the data can be generated from a known function $y = y(x)$. We assume the data points are equally spaced along the x-axis so that $x_{i+1} - x_i = h$ is a constant for $i = 0,1,2,...,n-1$. In this chapter we develop ways to approximate the derivatives of $y = y(x)$ given only the data points. We also develop ways to integrate the function $y = y(x)$ based solely upon the data points given.

19.1 NUMERICAL DIFFERENTIATION

The general method for deriving the numerical differentiation formulae is to differentiate the interpolating polynomial. Here we illustrate the derivation with Newton's forward difference formula only. The method of derivation being the same with regard to the other formulae. Consider Newton'forward difference formula:

$$y = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 + ... \tag{19.1}$$

where

$$x = x_0 + uh \tag{19.2}$$

Then

$$\frac{dy}{dx} = \frac{dy}{du}\cdot\frac{du}{dx} = \frac{1}{h}\left[\Delta y_0 + \frac{2u-1}{2}\Delta^2 y_0 + \frac{3u^2-6u+2}{6}\Delta^3 y_0 + ...\right]$$

This formula is for nontabular values of x. For tabular values of x, by setting $x = x_0$ we obtain $u = 0$ from 19.2. Hence equation 19.3 gives

$$\left[\frac{dy}{dx}\right]_{x=x_0} = \frac{1}{h}\left[\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \ldots\right] \tag{19.3}$$

Again for second order derivative, differentiating 19.3 further, we obtain

$$\frac{d^2y}{dx^2} = \frac{1}{h^2}\left[\Delta^2 y_0 + \frac{6u-6}{6}\Delta^3 y_0 + \frac{12u^2-36u+22}{24}\Delta^4 y_0 + \ldots\right] \tag{19.4}$$

from which, we get

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_0} = \frac{1}{h^2}\left[\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 - \frac{5}{6}\Delta^5 y_n + \ldots\right] \tag{19.5}$$

Similarly other higher order derivatives may be obtained by successive differentiation.

In a similar way, different formulae can be derived by starting with other interpolation formulae. Thus Newton's backward interpolation formula gives

$$\left[\frac{dy}{dx}\right]_{x=x_n} = \frac{1}{h}\left[\nabla y_n + \frac{1}{2}\nabla^2 y_n + \frac{1}{3}\nabla^3 y_n + \frac{1}{4}\nabla^4 y_n + \ldots\right] \tag{19.6}$$

and

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_n} = \frac{1}{h^2}\left[\nabla^2 y_n + \nabla^3 y_n + \frac{11}{12}\nabla^4 y_n + \frac{5}{6}\nabla^5 y_n + \ldots\right] \tag{19.7}$$

■ EXAMPLE 19.1

From the following table of values of x and y, obtain $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ for $x = 1.2$

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

The difference table is

x	y	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
1.0	2.7183						
		0.6018					
1.2	**3.3201**		0.1333				
		0.7351		0.0294			
1.4	4.0552		**0.1627**		0.0067		
		0.8978		**0.0361**		0.0013	
1.6	4.9530		0.1988		**0.0080**		0.0001
		1.0966		0.0441		**0.0014**	
1.8	6.0496		0.2429		0.0094		
		1.3395		0.0535			
2.0	7.3891		0.2964				
		1.6359					
2.2	9.0250						

Here $x_0 = 1.2$, $y_0 = 3.3201$ and $h = 0.2$ and

$$\left[\frac{dy}{dx}\right]_{x=x_0} = \frac{1}{h}\left[\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \ldots\right]$$

This gives

$$\begin{aligned}\left[\frac{dy}{dx}\right]_{x=1.2} &= \frac{1}{0.2}\left[0.7351 - \frac{1}{2}(0.1627) + \frac{1}{3}(0.0361) - \frac{1}{4}(0.0080) + \frac{1}{5}(0.0014)\right]\\ &= 3.3205\end{aligned}$$

and

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_0} = \frac{1}{h^2}\left[\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 - \frac{5}{6}\Delta^5 y_n + \ldots\right]$$

gives

$$\begin{aligned}\left[\frac{d^2y}{dx^2}\right]_{x=1.2} &= \frac{1}{0.04}\left[0.1627 - 0.0361 + \frac{11}{12}(0.0080) - \frac{5}{6}(0.0014)\right]\\ &= 3.318\end{aligned}$$

Problems

19.1 Compute $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ for $x = 1$, using following table:

x	1	2	3	4	5	6
y	1	8	27	64	125	216

19.2 Find $\frac{dy}{dx}$ at $x = 0.1$ from the following table :

x	0.1	0.2	0.3	0.4
y	0.9975	0.9900	0.9776	0.9604

19.3 From the following table of values of x and y, obtain $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ for $x = 1.2, 2.2$ and 1.6.

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	90.250

19.4 A rod is rotating in a plane. The following table gives the angle θ (in radians) through which the rod has turned for various values of time t (in seconds)

t	0	0.2	0.4	0.6	0.8	1.0	1.2
θ	0	0.12	0.49	1.12	2.02	3.20	4.67

Calculate the angular velocity and angular acceleration of the rod at $t = 0.6$ sec.

19.5 The table below gives the results of an observation. θ is the observed temperature in degrees centigrade of a vessel of cooling water, t is the time in minutes from the beginning of observations :

t	1	3	5	7	9
θ	85.3	74.5	67.0	60.5	54.3

Find the approximate rate of cooling at $t = 3$ and 3.5.

19.6 Find x for which y is maximum and find this value of y

x	1.2	1.3	1.4	1.5	1.6
y	0.9320	0.9636	0.9855	0.9975	0.9996

19.7 Find $f'''(5)$ from the data given below:

x	2	4	9	13	16	21	29
$f(x)$	57	1345	66340	402052	1118209	4287844	21242820

19.8 Find the numerical value of $y'(10^o)$ for $y = \sin x$ given that

$$\sin 0^o = 0.000, \sin 10^o = 0.1736, \sin 20^o = 0.3420, \sin 30^o = 0.5000, \sin 40^o = 0.6428$$

19.9 Find $\frac{d}{dx}(J_0)$ at $x = 0.1$ from the following table

x	0.0	0.1	0.2	0.3	0.4
$J_0(x)$	1	0.9975	0.99	0.9776	0.9604

19.10 Find the first and second derivatives for the function tabulated below at the point $x = 3.0$

x	3	3.2	3.4	3.6	3.8	4.0
y	-14	-10.032	-5.296	-0.256	6.672	14

19.11 A slider in a machine moves along a fixed straight rod. Its distance $x$cm along the rod is given below for various values of the time t seconds. Find the velocity of the slider and its acceleration when $t = 0.3$ second

t	0	0.1	0.2	0.3	0.4	0.5	0.6
x	30.13	31.62	32.87	33.64	33.95	33.81	33.24

19.12 From the table below, for what value of x, y is minimum ? Also find this value of y

x	3	4	5	6	7	8
y	0.205	0.240	0.259	0.262	0.250	0.224

19.13 Find the minimum value of y from the following table :

t	0.2	0.3	0.4	0.5	0.6	0.7
x	0.9182	0.8975	0.8873	0.8862	0.8935	0.9086

19.14 Given the following table :

x:	1	1.05	1.1	1.15	1.2	1.25	1.3
$f(x)=\sqrt{x}$:	1	1.0247	1.04881	1.07238	1.09544	1.11803	1.14014

Apply the above results to find $f'(1), f''(1)$ and $f'''(1)$

19.15 The following table gives values of pressure P and specific volume V of saturated steam :

P:	105	42.7	25.3	16.7	13
V:	2	4	6	8	10

Find
(a) the rate of change of pressure w.r.t. volume at $V = 2$
(b) the rate of change of volume w.r.t. pressure at $P = 105$.

19.16 y is a function of x satisfying the equation $xy'' + ay' + (x - b)y = 0$, where a and b are integers. Find the values of constants a and b if y is given by the following table:

x:	0.8	1	1.2	1.4	1.6	1.8	2	2.2
y:	1.73036	1.95532	2.19756	2.45693	2.73309	3.02549	2.3333	3.65563

19.17 Prove that

$$\frac{d}{dx}(y_x) = \frac{1}{h}(y_{x+h} - y_{x-h}) - \frac{1}{2h}(y_{x+2h} - y_{x-2h}) + \frac{1}{3h}(y_{x+3h} - y_{x-3h}) - \dots$$

19.2 NUMERICAL INTEGRATION

The process of computing $\int_a^b y dx$, where $y = f(x)$ is given by a set of tabulated values $[x_i, y_i]$ where $i = 0, 1, 2, \ldots, n$, $a = x_0$ and $b = x_n$ is called *numerical integration*. Since $y = f(x)$ is a single variable function, the process in general is known as *quadrature*.

19.2.1 Newton's Cotes Quadrature Formula

Let $I = \displaystyle\int_a^b y dx$, where $y = f(x)$ takes the values $y_0, y_1, y_2, \ldots, y_n$ for $x_0, x_1, x_2, \ldots, x_n$. Let us divide the interval (a, b) into n equal parts of width h, so that

$$a = x_0,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h,\ \ldots,\ x_n = x_0 + nh$$

Then

$$I = \int_{x_0}^{x_0+nh} y dx$$

On substituting $x = x_0 + rh$, so that $dx = hdr$, we get

$$\begin{aligned} I &= \int_{x_0}^{x_0+nh} y_r h dr \\ &= h \int_0^n \left[y_0 + r\Delta y_0 + \frac{r(r-1)}{2!}\Delta^2 y_0 + \frac{r(r-1)(r-2)}{3!}\Delta^3 y_0 + \ldots \right] \end{aligned}$$

Now on integrating term by term, we get

$$I = h\left[ny_0 + \frac{n^2}{2}\Delta y_0 + \frac{1}{2}\left\{\frac{n^3}{3} - \frac{n^2}{2}\right\}\Delta^2 y_0 + \frac{1}{6}\left\{\frac{n^4}{4} - n^3 - n^2\right\}\Delta^3 y_0 + ...\right] \tag{19.8}$$

Equation 19.8 is known as Newton Quadrature formula, which is also called as general qudrature formula.

19.3 TRAPEZODIAL RULE

Setting $n = 1$ in 19.8, we obtain

$$\int_{x_0}^{x_1} ydx = h\left(y_0 + \frac{1}{2}\Delta y_0\right) = \frac{h}{2}(y_0 + y_1)$$

For susequent intervals, similarly,

$$\int_{x_1}^{x_2} ydx = \frac{h}{2}(y_1 + y_2)$$
$$\int_{x_2}^{x_3} ydx = \frac{h}{2}(y_2 + y_3)$$
$$\cdots$$
$$\int_{x_{n-1}}^{x_n} ydx = \frac{h}{2}(y_{n-1} + y_n)$$

On adding all the above results,

$$\int_{x_0}^{x_n} ydx = \frac{h}{2}(y_0 + 2(y_1 + y_2 + ... + y_{n-1}) + y_n) \tag{19.9}$$

which is known as trepezodial rule.

■ **EXAMPLE 19.2**

find from the following table, the area bounded by the curve and x-axis from $x = 7.47$ to $x = 7.52$

x	7.47	7.48	7.49	7.50	7.51	7.52
$f(x)$	1.93	1.95	1.98	2.01	2.03	2.06

Solution: We know that the required area under curve would be

$$Area = \int_{7.47}^{7.52} f(x)dx$$

With $h = 0.01$, The trepezodial rule

$$\int_{x_0}^{x_n} ydx = \frac{h}{2}(y_0 + 2(y_1 + y_2 + ... + y_{n-1}) + y_n)$$

gives

$$Area = \frac{0.01}{2}[1.93 + 2(1.95 + 1.98 + 2.01 + 2.03) + 2.06]$$
$$= \frac{0.01}{2}[19.93] = 0.0996$$

19.4 SIMPSON'S 1/3 RULE

Setting $n = 2$ in 19.8, we obtain

$$\int_{x_0}^{x_2} y dx = h\left(y_0 + \Delta y_0 + \frac{1}{6}\Delta^2 y_0\right) = \frac{h}{3}(y_0 + 4y_1 + y_2)$$

For susequent intervals, similarly,

$$\int_{x_2}^{x_4} y dx = \frac{h}{3}(y_2 + 4y_3 + y_4)$$
$$\int_{x_4}^{x_6} y dx = \frac{h}{3}(y_4 + 4y_5 + y_6)$$
$$\ldots$$
$$\int_{x_{n-2}}^{x_n} y dx = \frac{h}{3}(y_{n-2} + 4y_{n-1} + y_n)$$

On adding all the above results,

$$\int_{x_0}^{x_n} y dx = \frac{h}{3}(y_0 + 4(y_1 + y_3 + \ldots + y_{n-1}) + 2(y_2 + y_4 + \ldots + y_{n-2}) + y_n) \qquad (19.10)$$

which is known as Simpson's 1/3 rule. This rule requires an even number of subintervals of width h.

■ EXAMPLE 19.3

The velocity v (metre/sec) of a particle at distance x (metre) is given in tabulated form:

x	0	10	20	30	40	50	60
v	47	58	65	64	58	55	35

Using simpson's 1/3 rule, find the time taken by the particle to travel 60 metre.

Solution: We know that

$$v = \frac{dx}{dt}$$

Therefore

$$t = \int \frac{1}{v} dx$$

Hence the time taken to travel 60 metre is

$$T = \int_0^{60} \frac{1}{v} dx$$

Since given data in table is for v, so now we preapre the table for $1/v$, because, here $f(x) = 1/v$. Therefore

x	0	10	20	30	40	50	60
$1/v$	0.02128	0.01724	0.01538	0.01562	0.01724	0.01818	0.02857

Now Simpson's 1/3 rule gives,

$$\int_{x_0}^{x_n} y dx = \frac{h}{3}(y_0 + 4(y_1 + y_3 + \ldots + y_{n-1}) + 2(y_2 + y_4 + \ldots + y_{n-2}) + y_n)$$
$$= \frac{10}{3}(0.02128 + 4(0.01724 + 0.01562 + 0.01818) + 2(0.01538 + 0.01724) + 0.02857)$$
$$= 1.064 \text{ sec}$$

Hence the time taken to travel 60 metre is 1.064 sec.

19.5 SIMPSON'S 3/8 RULE

Setting $n = 3$ in 19.8, we obtain

$$\int_{x_0}^{x_3} y dx = 3h\left(y_0 + \frac{3}{2}\Delta y_0 + \frac{3}{4}\Delta^2 y_0 + \frac{1}{8}\Delta^3 y_0\right) = \frac{3h}{8}(y_0 + 3y_1 + 3y_2 + y_3)$$

For susequent intervals, similarly,

$$\int_{x_3}^{x_6} y dx = \frac{3h}{8}(y_3 + 3y_4 + 3y_5 + y_6)$$
$$\int_{x_6}^{x_9} y dx = \frac{3h}{8}(y_6 + 3y_7 + 3y_8 + y_9)$$
$$\ldots$$
$$\int_{x_{n-3}}^{x_n} y dx = \frac{3h}{8}(y_{n-3} + 3y_{n-2} + 3y_{n-1} + y_n)$$

On adding all the above results,

$$\int_{x_0}^{x_n} y dx = \frac{3h}{8}(y_0 + 3(y_1 + y_2 + y_4 + y_5 + \ldots + y_{n-2} + y_{n-1}) + 2(y_3 + y_6 + \ldots + y_{n-3}) + y_n) \quad (19.11)$$

which is known as Simpson's 3/8 rule.

■ EXAMPLE 19.4

Evaluate $I = \int_0^6 \frac{1}{1+x} dx$ using Simpson's 3/8 rule.

Solution: Here we have

$$y = f(x) = \frac{1}{1+x}$$

where $xin[0,6]$ Consider 6 intervals, thus

$$h = \frac{6-0}{6} = 1$$

By using this information, we may prepare the following table:

x	0	1	2	3	4	5	6
y	1.00000	0.50000	0.33333	0.25000	0.20000	0.16667	0.14286

By Simpson rule (for 7 values)

$$\int_{x_0}^{x_n} ydx = \frac{3h}{8}(y_0 + 3(y_1 + y_2 + y_4 + y_5) + 2(y_3))$$
$$= \frac{3(1)}{8}(1.00000 + 3(0.50000 + 0.33333 + 0.20000 + 0.16667) + 2(0.25000) + 0.14286)$$
$$= 1.96607$$

Problems

19.18 Use Trapezoidal rule to evaluate $\int_0^1 x^3 ds$

19.19 Evaluate $\int_0^1 \frac{dx}{1+x^2}$

a) Simpson's $\frac{1}{3}$ rule taking $h = \frac{1}{4}$

b) Simpson's $\frac{3}{8}$ rule taking $h = \frac{1}{6}$

19.20 Evaluate $\int_0^6 \frac{dx}{1+x^2}$ by using

a) Simpson's one-third rule

b) Simposon's three-eight rule

c) Trapezoidal rule

19.21 Evaluate $\int_0^2 ydx$, where y is given by the following table:

x	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
y	1.23	1.58	2.03	4.32	6.25	8.36	10.23	12.45

19.22 Find $\int_1^{11} f(x)dx$, where $f(x)$ is given by the following table, using a suitable integration formula.

x	1	2	3	4	5	6	7	8	9	10	11
$f(x)$	543	512	501	489	453	400	352	310	250	172	95

19.23 Evaluate $\int_0^1 \frac{dx}{1+x}$ by dividing the interval of integration into 8 equal parts. Hence find $\ln 2$ approximately.

19.24 Evaluate $\int_1^2 e^{x/2} dx$ using four intervals.

19.25 Find $\int_0^6 \frac{e^x}{1+x} dx$ approximately using Simpson's $\frac{3}{8}$th rule of integration.

19.26 A train is moving at the speed of 30 m/sec Suddenly brakes are applied. The speed of the train per second after t seconds is given by

Time(t)	0	5	10	15	20	25	30	35	40	45
Speed (v)	30	24	19	16	13	11	10	8	7	5

Apply simpson's three-eighth rule to determine the distance moved by the train in 45 seconds.

19.27 A river is 80 m wide. The depth y of the river at a distance x from one bank is given by the following table :

x	0	10	20	30	40	50	60	70	80
y	0	4	7	9	12	15	14	8	3

Find the approximate area of cross-section of the river using

a) Trapezodial rule.

b) Simplosn's $\frac{1}{3}$rd rule.

19.28 Evaluate $\int_{0.2}^{1.4}(\sin x - \ln x + e^x)dx$ approximately using Simpson's rules correct to 4 decimals places.

19.29 A solid of revolution is formed by rotating about x-axis, the lines $x = 0$ and $x = 1$ and a curve through the points with the following coordinates.

x	0	0.25	0.5	0.75	1
y	1	0.9896	0.9589	0.9089	0.8415

Estimate the volume of the solid formed using Simpson's rule.

19.30 Using the following table of values, approximate by Simpson's rule, the arc length of the graph $y = \frac{1}{x}$ between the points $(1,1)$ and $(5, \frac{1}{5})$

x	1	2	3	4	5
$\sqrt{\frac{1+x^4}{x^4}}$	1.414	1.031	1.007	1.002	1.001

19.31 From the following values of $y = f(x)$ in the given range of values of x, find the position of the centroid of the area under the curve and the x-axis

x	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
y	1	4	8	4	1

Also find

a) the volume of sokid obtained by revolving the above area about x-axis.

b) the moment of inertia of the area about x-axis.

19.32 A reservoir discharging water through sluices at a depth h below the water surface, has a surface area A for various values of h as given below:

h (in meters)	10	11	12	13	14
A (in sq. metres)	950	1070	1200	1350	1530

Here t denotes time in minutes, the rate of fall of the surface is given by $\frac{dh}{dt} = -48A\sqrt{h}$ Estimate the time taken for the water level to fall from 14 to 10 m above the sluices.

19.33 Evaluate $\int_1^2 \frac{1}{x}dx$ by Simpson's $\frac{1}{3}$rd rule with four strips and determine the error by direct integration.

19.34 Evaluate the integral $\int_0^{\frac{\pi}{2}} \sqrt{\cos\theta}d\theta$ by dividing the interval into 6 parts.

19.35 Evaluate $\int_4^{5.2} \ln xdx$ by Trapezoidal rule and Simpson's $\frac{3}{8}$th rule. Also write its programme in 'C' language.

19.36 Evaluate $\int_{30}^{90} log(\sin x)dx$ by Simpson's $1/3$rd rule by dividing the interval into 6 parts.

19.37 Evaluate using Trapezoidal rule

a) $\int_0^{\pi} t\sin tdt$

b) $\int_{-2}^{2} \frac{tdt}{5+2t}$

19.38 Evaluate $\int_3^7 x^2 \log xdx$ taking 4 strips.

19.39 Find an approximate value of $\ln 5$ by calculating to 4 decimal places, by Simpson's $\frac{1}{3}$rd rule, $\int_0^5 \frac{dx}{4x+5}$ dividing the range into 10 equal parts.

19.40 The velocity of a train which starts from rest is given by the following table, time being reckoned in minutes from the start and speed in kilometres per hour :

Minutes	0	2	4	6	8	10	12	14	16	18	20
Speed (km/hr.)	0	10	18	25	29	32	20	117 5	2	0	

Estimate the total distance in 20 minutes. Hint. Here step-size $h = \frac{2}{60}$

19.41 A rocket is launched from the ground. Its acceleration is registered during the first 80 seconds and is given in the table below. Using Simpson's $\frac{1}{3}$rd rule, find the velocity of the rocket at $t = 80$ seconds.

t (sec)	0	10	20	30	40	50	60	70	80
f(cm/sec^2)	30	31.63	33.34	35.47	37.75	40.33	43.25	46.69	50.67

19.42 A curve is drawn to pass through the points given by the following table :

x	1	1.5	2	2.5	3	3.5	4
y	2	2.4	2.7	2.8	3	2.6	2.1

Find

a) Centre of gravity of the area.

b) Volume of the solid of revolution.

c) The area bounded by the curve, the x-axis and lines $x = 1, x = 4$.

19.43 In an experiment, a quantity G was measured as follows $G(20) = 95.9, G(21) = 96.85, G(22) = 97.77G(23)98.68, G(24) = 99.56, G(25) = 100.41, G(26) = 101.24$. Compute $\int_{20}^{26} G(x)dx$ by Simpson's rules.

19.44 Using the data of the following table, compute the integral $\int_{0.5}^{1.1} xydx$ by Simpson's rule

x	0.5	0.6	0.7	0.8	0.9	1.0	1.1
y	0.4804	0.5669	0.6490	0.7262	0.7985	0.8658	0.9281

19.45 Find the value of $\ln 3$ from $\int_0^1 \frac{x^2}{1+x^3}dx$ using Simpson's $\frac{1}{3}$rd rule by dividing the range of integration into four equal parts. Also estimate the error.

19.46 The function $f(x)$ is known at one point x' in the interval $[a,b]$. Using this value, $f(x)$ can be expressed as $f(x)p_0(x) + f'\xi(x)(x-x')$ for $x \in (a,b)$ where $p_0(x)$ is the zeroth-order interpolating polynomial $p_0(x) = f(x')$ and $\xi(x) \in (a,b)$. Integrate this expression from a to b to derive a quadrature rule with error term. Simplify the error term for the case when $x' = a$.

CHAPTER 20

ORDINARY DIFFERENTIAL EQUATIONS

Many ordinary differential equations encountered do not have easily obtainable closed form solutions, and we must seek other methods by which solutions can be constructed. Numerical methods provide an alternative way of constructing solutions to these sometimes difficult problems. In this chapter we present an introduction to some numerical methods which can be applied to a wide variety of ordinary differential equations. These methods can be programmed into a digital computer or even programmed into some hand-held calculators. Many of the numerical techniques introduced in this chapter are readily available in the form of subroutine packages available from the internet.

20.1 PICARD'S METHOD

20.1.1 First Order Differential Equations

Consider the first order differential equation

$$\frac{dy}{dx} = f(x,y) \tag{20.1}$$

subject to $y(x = x_0) = y_0$. On integrating 20.1 with respect to x between the limits (x_0, x), we get

$$\int_{y_0}^{y} dy = \int_{x_0}^{x} f(x,y)dx$$
$$y - y_0 = \int_{x_0}^{x} f(x,y)dx$$

$$y = y_0 + \int_{x_0}^{x} f(x,y)dx$$

In Picard's method, to obtain nth successive approximation, we write above equations as

$$y_n = y_0 + \int_{x_0}^{x} f(x,y_{n-1})dx \tag{20.2}$$

To terminate the process, we compare the values y_n and y_{n-1} up to desired accuracy.

EXAMPLE 20.1

Use Picard's method to approximate the value of y at $x = 0.1$ given that $y(0) = 1$ and $\dfrac{dy}{dx} = \dfrac{y-x}{y+x}$

Given $\dfrac{dy}{dx} = \dfrac{y-x}{y+x}$ Hence Picard's Iteration equation is

$$y_n = y_0 + \int_{x_0}^{x} \frac{y_{n-1}-x}{y_{n-1}+x} dx$$

which gives First approximation

$$\begin{aligned} y_1 &= y_0 + \int_{x_0}^{x} \frac{y_0 - x}{y_0 + x} dx \\ &= 1 + \int_{x_0}^{x} \frac{1-x}{1+x} dx = 1 + \int_{x_0}^{x} \left[\frac{2}{1+x} - \right] dx \\ &= 1 + (2\log(1+x) - x)|_0^x = 1 + 2\log(1+x) - x \end{aligned}$$

Second approximation

$$\begin{aligned} y_2 &= y_1 + \int_{x_0}^{x} \frac{y_1 - x}{y_1 + x} dx \\ &= 1 + \int_{x_0}^{x} \frac{1 + 2\log(1+x) - x - x}{1 + 2\log(1+x) - x + x} dx \\ &= 1 + 2\int_{x_0}^{x} \frac{x}{1 + 2\log(1+x)} dx \end{aligned}$$

which is quite difficult to integrate. So we stop our process here.

We use only first order approximation. By putting $x = 0.1$

$$y_1 = 1 + (2\log(1+0.1) - 0.1) = 0.9828$$

20.1.2 Simultaneous First Order Differential Equations

Consider the simultaneous first order differential equation

$$\begin{aligned} \frac{dy}{dx} &= \phi(x,y,z) \\ \frac{dz}{dx} &= \psi(x,y,z) \end{aligned}$$

Piacrd's successive approximation for these equation is

$$y_n = y_0 + \int_{x_0}^{x} \phi(x, y_{n-1}, z_{n-1})dx$$
$$z_n = z_0 + \int_{x_0}^{x} \psi(x, y_{n-1}, z_{n-1})dx$$

Using these equations, we may iterate the process up to desired accuracy of y and z.

EXAMPLE 20.2

Approximate y and z at $x = 0.1$ using the Picard's method to solve the equations

$$\frac{dy}{dx} = z$$
$$\frac{dz}{dx} = x^3(y+z)$$

given that $y(0) = 1$ and $z(0) = 0.5$.

In Picrad's Method

$$y_n = y_0 + \int_{x_0}^{x} \phi(x, y_{n-1}, z_{n-1})dx$$
$$z_n = z_0 + \int_{x_0}^{x} \psi(x, y_{n-1}, z_{n-1})dx$$

Given $x_0 = 0$, $y_0 = 1$, and $z_0 = 0.5$. For First Approximation

$$y_1 = y_0 + \int_{x_0}^{x} \phi(x, y_0, z_0)dx = 1 + \int_0^x \frac{1}{2}dx = 1 + \frac{1}{2}x$$
$$z_1 = z_0 + \int_{x_0}^{x} \psi(x, y_0, z_0)dx = \frac{1}{2} + \int_0^x x^3(1 + \frac{1}{2})dx = \frac{1}{2} + \frac{3x^4}{8}$$

For Second Approximation

$$y_2 = y_0 + \int_{x_0}^{x} \phi(x, y_1, z_1)dx = 1 + \int_0^x (\frac{1}{2} + \frac{3}{8}x^4)dx = 1 + \frac{1}{2}x + \frac{3x^5}{40}$$
$$z_2 = z_0 + \int_{x_0}^{x} \psi(x, y_1, z_1)dx = \frac{1}{2} + \int_0^x x^3(1 + \frac{x}{2} + \frac{1}{2} + \frac{3x^4}{8})dx = \frac{1}{2} + \frac{3x^4}{8} + \frac{x^5}{10} + \frac{3x^8}{64}$$

Now when $x = 0.1$ $y_1 = 1.05$ and $z = 0.5000375$ and $y_2 = 1.0500008$ and $z = 0.5000385$.

Hence $y(0.1) = 1.0500$ and $z(0.1) = 0.5000$. More accuracy in approximations needs more iterations.

20.1.3 Second Order Differential Equations

Consider the second order differential equation

$$\frac{d^2y}{dx^2} = f(x, y, \frac{dy}{dx}$$

By putting $\dfrac{dy}{dx} = z$, the above equation reduces to first order simultaneous differential equations:

$$\frac{dy}{dx} = z$$
$$\frac{dz}{dx} = f(x,y,z)$$

Now these equation can be solved as discussed in previous section.

■ EXAMPLE 20.3

Use Picard's method to approximate y when $x = 0$ for the differential equation

$$\frac{d^2y}{dx^2} + 2x\frac{dy}{dx} + y = 0$$

given that $y(0) = 0.5$ and $\left[\dfrac{dy}{dx}\right]_{x=0} = 0.1$.

Let $ds\frac{dy}{dx} = z$. Thus $\dfrac{d^y}{dx^2} = \dfrac{dz}{dx}$. On substituting thse results,

$$\frac{d^2y}{dx^2} + 2x\frac{dy}{dx} + y = 0$$

reduces to

$$\frac{dz}{dx} = -(2xz + y)$$

Thus, now proble changes to

$$\frac{dy}{dx} = z$$
$$\frac{dz}{dx} = -(2xz + y)$$

with conditions $y(0) = 0.5$ and $\left[\dfrac{dy}{dx}\right]_{x=0} = 0.1$.

First Approximation

$$\begin{aligned} y_1 &= 0.5 + \int_0^x z_0 dx = 0.5 + \int_0^x (0.1)dx \\ &= 0.5 + (0.1)x \\ z_1 &= 0.1 - \int_0^x (2xz_0 + y_0)dx = 0.1 - \int_0^x ((0.2)x + 0.5)dx \\ &= 0.1 - (0.5)x + (0.1)x^2 \end{aligned}$$

Second Approximation

$$\begin{aligned} y_2 &= 0.5 + \int_0^x z_1 dx = 0.5 + \int_0^x (0.1 - (0.5)x + (0.1)x^2)dx \\ &= 0.5 + (0.1)x - \frac{x^2}{4} - \frac{x^3}{30} \\ z_2 &= 0.1 - \int_0^x (2xz_1 + y_1)dx = 0.1 - \int_0^x (2x(0.1 - (0.5)x + (0.1)x^2) + (0.5 + (0.1)x))dx \\ &= 0.1 - (0.5)x - \frac{3x^2}{20} + \frac{x^3}{3} + \frac{x^4}{20} \end{aligned}$$

similarly, we may compute

$$y_3 = 0.5 + (0.1)x - \frac{x^2}{4} - \frac{x^3}{20} + \frac{x^4}{12} + \frac{x^5}{100}$$
$$z_3 = 0.1 - (0.5)x - \frac{3x^2}{20} - \frac{5x^3}{12} + (0.2)x^4 + \frac{2x^5}{15} + \frac{x^6}{60}$$

Now at $x = 0.1$,
$y_1 = 0.51$, $y_2 = 0.50746$, and $y_3 = 0.50745$
Hence, $y(0.1) = 0.5075$ correct upto four decimal places.

Problems

20.1 Given the differential equation

$$\frac{dy}{dx} = \frac{x^2}{y^2+1}$$

with the initial condition $y = 0$ when $x = 0$. Use Picard's method to obtain y for $x = 0.25, 0.5$ and 1.0 correct to three decimal places.

20.2 Use Picard's method to obtain y for $x = 0.2$. Given :

$$\frac{dy}{dx} = x - y$$

with initial condition $y = 1$ when $x = 0$

20.3 If $\dfrac{dy}{dx} = \dfrac{y-x}{y+x}$, find the value of y at x=0.1 using Picard's method. Given that $y_0 = 1$.

20.4 Obtain y when $x = 0.1, x = 0.2$. Given that $\dfrac{dy}{dx} = x + y$; $y(0) = 1$. Check the result with exact value.

20.5 Find the solution of $\dfrac{dy}{dx} = 1 + xy, y(0) = 1$ which passes through $(0,1)$ in the interval $(0,0.5)$ such that the value of y is correct to three decimal places (use the whole interval as one interval only). Take $h = 0.1$.

20.6 Approximate y and z by using Picard's method for the particular solution of

$$ds\frac{dy}{dx} = x + z$$

$$\frac{dz}{dx} = x - y^2$$

given that $y = 2, z = 1$ when $x = 0$.

20.7 Solve by Picard's method, the differential equations

$$\frac{dy}{dx} = z, \ \frac{dz}{dx} = x^3(y+z)$$

where $y = 1, z = 0.5$ at $x = 0$. Obtain the values of y and z from III approximation when $x = 0.2$ and $x = 0.5$.

20.8 Find $y(0,2)$ if $\dfrac{dy}{dx} = log(x+y)$; $y(0) = 1$. Use Picard's method.

20.9 Employ Picard's method to obtain the solution of

$$\frac{dy}{dx} = x^2 + y^2$$

for $x = 0.1$ correct to four places of decimal given that $y = 0$ when $x = 0$.

20.10 Find an approximate value of y when $x = 0.1$ if $\frac{dy}{dx} = x - y^2$ and $y = 1$ at $x = 0$ using Picard's method.

20.11 Approximate y and z by using Picard's method for the solution of simultaneous differential equations

$$\frac{dy}{dx} = 2x + z$$

,

$$\frac{dz}{dx} = 3xy + x^2 z$$

with $y = 2, z = 0$ at $x = 0$ upto third approximation.

20.12 Using Picard's method, obtain the solution of

$$\frac{dy}{dx} = x(1 + x^3 y), y(0) = 3$$

Tabulate the values of $y(0.1), y(0.2)$.

20.2 EULER'S METHOD

20.2.1 First Order Differential Equations

Consider the differential equation

$$\frac{dy}{dx} = f(x, y) \tag{20.3}$$

where $y(x_0) = y_0$.

Suppose that we wish to find successively $y_1,\ y_2,\ \ldots,\ y_n$ where y_m is the value of y corresponding to $x = x_m$, where $x_m = x_0 + mh$, $m = 1,\ 2, \ldots, h$ being small. Here, we use the property that in a small interval, a curve is nearly a straight line.

Thus, in the interval x_0 to x_1 of x, we approximate the curve by the tangent at the point (x_0, y_0).

Therefore, the equation of tangent at (x_0, y_0) is

$$y - y_0 = \left(\frac{dy}{dx}\right)_{(x_0, y_0)} (x - x_0)$$

$$= f(x_0, y_0)(x - x_0) \qquad \text{[From equation 20.3]}$$

$$\text{or} \qquad y = y_0 + (x - x_0) f(x_0, y_0)$$

Hence, the value of y corresponding to $x = x_1$ is

$$y_1 = y_0 + (x_1 - x_0) f(x_0, y_0)$$

or

$$y_1 = y_0 + h f(x_0, y_0) \tag{20.4}$$

Since the curve is approximated by the tangent in $[x_0, x_1]$, Equation 20.4 gives the approximated value of y_1.

Simialrly, approximating the curve int he next interval $[x_1, x_2]$ by a line through (x_1, y_1) with slope $f(x_1, y_1)$, we get

$$y_2 = y_1 + hf(x_1, y_1) \tag{20.5}$$

Proceeding on, in general it can be shown that

$$y_{m+1} = y_m + hf(x_m, y_m) \tag{20.6}$$

EXAMPLE 20.4

Given $dy/dx = (y-x)/(y+x)$ with $y = 1$ at $x = 0$. Find y for $x = 0.1$ in five steps.

Solution: Here, we have

$$f(x,y) = \frac{dy}{dx} = \frac{y-x}{y+x},$$

with $x_0 = 0, y_0 = 1$. Let $h = \frac{0.1}{5} = 0.020$. Then by Euler's method

$$\begin{aligned} y_1 &= y_0 + hf(x_0, y_0) \\ &= 1 + 0.20\left(\frac{1-0}{1+0}\right) = 1.02 \end{aligned}$$

Again

$$\begin{aligned} y_2 &= y_1 + hf(x_1, y_1), \quad x_1 = x_0 + h = 0 + 0.02 \\ &= 1.02 + 0.02 \left[\frac{1.02 - 0.02}{1.02 + 0.02}\right] = 1.0392 \end{aligned}$$

Now

$$\begin{aligned} y_3 &= y_2 + hf(x_2, y_2) \quad x_2 = x_1 + h = 0.04 \\ &= 1.0392 + 0.02 \left[\frac{1.0392 - 0.04}{1.0392 + 0.04}\right] = 1.0577 \end{aligned}$$

$$\begin{aligned} y_4 &= y_3 + hf(x_3, y_3) \quad x_3 = x_2 + h = 0.06 \\ &= 1.0577 + 0.02 \left[\frac{1.0577 - 0.06}{1.0577 + 0.06}\right] = 1.0756 \end{aligned}$$

$$\begin{aligned} y_5 &= y_4 + hf(x_4, y_4) \quad x_3 = x_2 + h = 0.06 \\ &= 1.0756 + 0.02 \left[\frac{1.0756 - 0.08}{1.0756 + 0.08}\right] = 1.0928 \end{aligned}$$

Therefore y at $x = 0.1$ is 1.0928 □

20.2.2 Simultaneous First Order Differential Equations

Euler's methed can be extended to the solution of system of ordinary differential equations. The folowing example illustrates the procedure.

EXAMPLE 20.5

Solve $\dfrac{dy}{dx} = -0.5y, \dfrac{dz}{dx} = 4 - 0.3z - 0.1y$ using Euler's method, assuming that $x = 0$, $y = 4$, and $z = 6$. Integrate to $x = 2$ with $h = 0.5$.

Solution:

We have $\dfrac{dy}{dx} = -0.5y$, $\dfrac{dz}{dx} = 4 - 0.3z - 0.1y$

Euler's formula is

$$y_{n+1} = y_n + hf(y_n, z_n)$$

Therefore

$y_1 = y_0 + hf(y_0, z_0) = 4 + 0.5[-0.5(4)] = 3$

$z_1 = z_0 + hf(y_0, z_0) = 6 + 0.5[4 - 0.3(6) - 0.1(4)] = 6.9$

$y_2 = y_1 + hf(y_1, z_1) = 3 + 0.5[-0.5(3)] = 2.25$

$z_2 = z_1 + hf(y_1, z_1) = 6.9 + 0.5[4 - 0.3(6.9) - 0.1(3)] = 7.715$

$y_3 = y_2 + hf(y_2, z_2) = 2.25 + 0.5[-0.5(2.25)] = 1.6875$

$z_3 = z_2 + hf(y_2, z_2) = 7.715 + 0.5[4 - 0.3(7.715) - 0.1(2.25)] = 8.44525$

$y_4 = y_3 + hf(y_3, h_3) = 1.6875 + 0.5[-0.5(1.6875)] = 1.265625$

$z_4 = z_3 + hf(y_3, z_3) = 8.44525 + 0.5[4 - 0.3(8.44525) - 0.1(1.6875)] = 9.0940875$ □

Problems

20.13 Apply Euler's method to the initial value problem $\dfrac{dy}{dx} = x + y, y(0) = 0$ $x = 0$ at $x = 1.0$ taking $h = 0.2$

20.14 Use Euler's method with $h = 0.1$ to solve the differential equation $\dfrac{dy}{dx} = x^2 + y^2$, $y(0) = 1$ in the range $x = 0$ to $x = 0.3$

20.15 Solve for y at $x = 1.05$ by Euler's method, the differential equation

$$\frac{dy}{dx} = 2 - \left(\frac{y}{x}\right)$$

where $y = 2$ when $x = 1$. (Take $h = 0.05$)

20.16 Use Euler's method to compute y for $x = .05$ and .10. Given that

$$\frac{dy}{dx} = x + y$$

with the initial condition $x_0 = 0, y_0 = 1$. Given the correct result upto 4 decimal places.

20.17 Using Euler's method, compute $y(0.04)$ for the differential equation $\dfrac{dy}{dx} = -y$; $y(0) = 1$. Take $h = 0.01$

20.18 Given $\dfrac{dy}{dx} = x^3 + y$, $y(0) = 1$. Compute $y(0.02$ by Euler's method taking $h = 0.01$.

20.19 Find $y(1)$ by Euler's method from the differential equation $\dfrac{dy}{dx} = \dfrac{-y}{1+x}$ when $y(0.3) = 2$. Convert upto four decimal places taking step length $h = 0.1$

20.3 RUNGE- KUTTA METHOD

C Runge and W. Kutta, two German mathematicians, developed this method and as such the method was named after them. It was developed to avoid the computation of higher order derivations which Taylor's method involves. Runge-Kutta formulas posses the advantage of requiring only the functional values at some selected points. This method agrees with Taylor's series solutions upto the term h^r where r differs from method to method and is called the order of that method. Here, we will discuss Runge-Kutta method of order four.

20.3.1 Fourth Order Runge-Kutta Method

Computational procedure for Runge-Kutta Method of Order Four . The computational procedure for Runge-Kutta method of order four is given as follows :

Step 1. Define $f(x,y), x_0, y_0$,and x_n.

Step 2. Find h by using $\frac{(x_n - x_0)}{n}$

step 3. Find k_1, k_2, k_3, k_4 and

$$
\begin{aligned}
k_1 &= hf(x_n, y_n),\\
k_2 &= hf\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}\right),\\
k_3 &= hf\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}\right),\\
k_4 &= hf(x_n + h, y_n + k_3),
\end{aligned}
$$

and therefore $k = \dfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$ Step 4. Find $y_{n+1} = y_n + k$., then do successive iteration up to required accuracy.

■ **EXAMPLE 20.6**

Use the Runge-Kutta fourth order method to find the value of y when $x = 1$ given that $y = 1$, when $x = 0$ (taking $n = 2$) and $\frac{dy}{dx} = \frac{(y-x)}{(y+x)}$.

Solution: We have

$$f(x,y) = \frac{y-x}{y+x},\ x_0 = 0,\ y_0 = 1,\ h = 0.5$$

Therefore

$$
\begin{aligned}
k_1 &= hf(x_0, y_0) = 0.5\left[\frac{1-0}{1+0}\right] = 0.50\\
k_2 &= hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = (0.5)\left(\frac{1.25-0.25}{1.25+0.25}\right) = 0.33333\\
k_3 &= hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = (0.5)\left(\frac{1.16666-0.25}{1.16666+0.25}\right) = 0.41176\\
k_4 &= hf(x_0 + h, y_0 + k_3) = (0.5)\left(\frac{1.41176-0.5}{1.41176+0.5}\right) = 0.23846
\end{aligned}
$$

Therefore

$$k = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$
$$= \frac{1}{6}(0.5 + 2 \times 0.33333 + 2 \times *0.41176 + 0.23846)$$
$$= 0.37144$$
$$\therefore y(0.5) = y_0 + k = 1 + 0.37144 = 1.37144$$

□

20.3.2 Simultaneous First Order Differential Equations

The simultaneous equations $\frac{dy}{dx} = f_1(x,y,z)$ and $\frac{dz}{dx} = f_2(x,y,z)$ given the initial conditions $y(x_0) = y_0,\ z(x_0) = z_0$.

[Here, x is independent variable while y and z are dependent.]

Now, starting from (x_0, y_0, z_0) the increments $\triangle y$ and $\triangle z$ in y and z respectively are given by formulae,

$$k_1 = hf_1(x_0, y_0, z_0)$$
$$l_1 = hf_2(x_0, y_0, z_0)$$
$$k_2 = hf_1(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1, z_0 + \frac{1}{2}l_1)$$
$$l_2 = hf_2\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1, z_0 + \frac{1}{2}l_1\right)$$
$$k_3 = hf_1(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_2, z_0 + \frac{1}{2}l_2)$$
$$l_3 = hf_2\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_2, z_0 + \frac{1}{2}l_2\right)$$

$$k_4 = hf_1(x_0 + h, y_0 + k_3, z_0 + l_3)$$
$$l_4 = hf_2(x_0 + h, y_0 + k_3, z_0 + l_3)$$

$$y_1 = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$
$$z_1 = z_0 + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4)$$

Thus, we get (x_1, y_1, z_1). Similarly we can get (x_2, y_2, z_2) by repeating the above method starting with (x_1, y_1, z_1).

■ **EXAMPLE 20.7**

Solve $\frac{dy}{dx} = yz + x$, $\frac{dz}{dx} = dz + y$, given that $y(0) = 1$, $z(0) = -1$ for $y(0.1)$, $z(0.1)$.

Solution: Here

$$f_1(x,y,z) = yz + x,\ f_2(x,y,z) = xz + y$$

$h = 0.1,\ x_0 = 0,\ y_0 = 1,\ z_0 = -1$

$$k_1 = hf_1(x_0, y_0, z_0) = h(y_0 z_0 + x_0) = -0.1$$
$$l_1 = hf_2(x_0, y_0, z_0) = h(x_0 z_0 + y_0) = 0.1$$

$$\begin{aligned} k_2 &= hf_1(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}, z_0 + \frac{l_1}{2}) \\ &= hf_1(0.05, 0.95, -0.95) = -0.08525 \\ l_2 &= hf_2(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}, z_0 + \frac{l_1}{2}) \\ &= hf_2(0.05, 0.95, -0.95) = 0.09025 \end{aligned}$$

$$\begin{aligned} k_3 &= hf_1(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}, z_0 + \frac{l_2}{2}) \\ &= hf_1(0.05, 0.957375, --0.945875) \\ &= -0.0864173 \\ l_3 &= hf_2(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}, z_0 + \frac{l_2}{2}) \\ &= hf_2(0.05, 0.957375, -0.954875) \\ &= -0.0864173 \end{aligned}$$

$$k_4 = hf_1(x_0 + h, y_0 + k_3, z_0 + l_3) = -0.073048$$
$$l_4 = hf_2(x_0 + h, y_0 + k_3, z_0 + l_3) = -0.0822679$$

Therefore $k = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = -0.0860637$

$$l = \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4) = 0.0907823$$

$$\therefore y_1 = y(0.1) = y_0 + k = 1 - 0.0860637 = 0.9139363$$
$$z_1 = z(0.1) = z_0 + k = -1 + 0.0907823 = -0.9092176$$

□

20.3.3 Second Order Differential Equations

Second order differentail are best treated by transforming the given equation into a system of first order simultaneous equations which can be solved by one of the aforesaid metheds.consider, for example the second order differential equation

$$\frac{d^2y}{dx^2} = f(x, y, \frac{dy}{dx});$$
$$y(x_0) = (\frac{dy}{dx}) = y_0$$

Substituing $\frac{dy}{dx} = z$, we get

$$\frac{dz}{dx} = f(x,y,z)$$

$$y(x_0) = y_0' \, z(x_0) = y_0$$

This constitute the equivalent system of simultaneous equations.

EXAMPLE 20.8

Use Runge Kutta methed to find $y(0.2)$ for the equation

$$\frac{d^2y}{dx^2} = x\frac{dy}{dx} - y$$

given that $y = 1, \frac{dy}{dx} = 0$, when $x = 0$

Solution:

Substituting $\dfrac{dy}{dx} = z = f(x,y,z)$. The given equation reduces to

$$\frac{d^2y}{dx^2} = xz - y = g(x,y,z)$$

The initial conditions are given by $x_0 = 0, y_0 = 1, z_0 = 0$, Also$h = 0.2$

$$k_1 = hf(x_0,y_0,z_0) = hz_0 = 0.2 \times 0 = 0$$
$$l_1 = hg(x_0,y_0,z_0) = h(x_0z_0 - y_0) = 0.2 \times (0-1) = -0.2$$

$$k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}, z_0 + \frac{l_1}{2}\right) = h\left(z_0 + \frac{l_1}{2}\right) = 0.2\left(0 - \frac{0.2}{2}\right) = -0.02$$
$$l_2 = hg\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}, z_0 + \frac{l_1}{2}\right) = h\left[\left(x_0 + \frac{h}{2}\right)\left(z_0 + \frac{l_1}{2}\right) - \left(y_0 + \frac{k_1}{2}\right)\right]$$
$$= 0.2\left[\left(0 + \frac{0.2}{2}\right)\left(0 - \frac{0.2}{2}\right) - \left(1 + \frac{0}{2}\right)\right] = -0.202$$

$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}, z_0 + \frac{l_2}{2}\right) = h\left(z_0 + \frac{l_2}{2}\right) = 0.2\left(0 - \frac{0.202}{2}\right) = -0.0202$$
$$l_3 = hg\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}, z_0 + \frac{l_2}{2}\right) = h\left[\left(x_0 + \frac{h}{2}\right)\left(z_0 + \frac{l_2}{2}\right) - \left(y_0 + \frac{k_2}{2}\right)\right]$$
$$= 0.2\left[\left(0 + \frac{0.2}{2}\right)\left(0 - \frac{0.202}{2}\right) - \left(0 - \frac{0.02}{2}\right)\right] = -0.20002$$

$$k_4 = hf(x_0 + h, y_0 + k_3, z_0 + l_3) = h(z_0 + l_3) = 0.2(0.20002) = -0.040004$$
$$l_4 = hg(x_0 + h, y_0 + k_3, z_0 + l_3) = h[(x_0 + h)(z_0 + l_3) - (y_0 + k_3)]$$
$$= 0.2[(0.2)(-0.20002) - (1 - 0.0202)] = -0.2039608$$

This given at $x = 0.2$

$$\begin{aligned} y(0.2) &= y(0) + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ &= 1 + \frac{1}{6}[0 + 2(-0.02) + 2(-0.0202) + (-0.040004)] \\ &= 0.97993267 \end{aligned}$$

$$\begin{aligned} z(0.2) &= z(0) + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4) \\ &= 0 + \frac{1}{6}[-0.2 + 2(-0.202) + 2(-0.20002) - 0.2039608] \\ &= -0.20133347 \end{aligned}$$

□

Problems

20.20 Solve the equation $\frac{dy}{dx} = x + y$ with initial condition $y(0) = 1$ by Runge-Kutta rule. from $x = 0$ to $x = 0.4$ with $h = 0.1$

20.21 Use Runge-Kutta method to approximate y when $x = 0.1$ given that $x = 0$ when $y = 1$ and $\frac{dy}{dx} = x + y$

20.22 Apply Runge-Kutta fourth order method to solver $10\frac{dy}{dx} = x^2 + y^2; y(0) = 1$ for $0 < x \leq 0.4$ and $h = 0.1$

20.23 Use Runge-Kutta Fourth order formula to find $y(1.4)$ if $y(1) = 2$ and $\frac{dy}{dx} = xy$. Take $h = 0.2$

20.24 Solve $y' = \frac{1}{x+y}, y(0) = 1$ for $x = 0.5$ to $x = 1$ by Runge-Kutta method $(h = 0.5)$

20.25 Using Runge-Kutta method of Fourth order, solve for $y(0.1), y(0.2)$ and $y(0.3)$ given that $y' = xy + y^2, y(0) = 1$

20.26 Solve $\frac{dy}{dx} = x + z$ $\frac{dz}{dx} = x - y^2$ for $y(0.1), z(0.1)$ given that $y(0) = 2, z(0) = 1$ by Runge-Kutta method.

20.27 Write the main steps to be followed in using the Runge-Kutta method of fourth order to solve an ordinary diff. equation fo first order. Hence solve $\frac{dy}{dx} = x^3 + y^3, y(0) = 1$ and step length $h = 0.1$ upto three iterations.

20.28 Given $\frac{dy}{dx} = xy$ with $y(1) = 5$. Using Fourth order Runge-Kutta method, find the solution in the interval $(1, 1.5)$ using step size $h = 0.1$

20.29 Using Runge-Kutta method of fourth order, solve the following differential equation : $\frac{dy}{dx} = \frac{y^2 - x^2}{y^2 + x^2}$ with $y(0) = 1$ at $x = 0.2, 0.4$ Also write computer program in 'C'

20.30 Solve $\frac{dy}{dx} = yz + x$, $\frac{dz}{dx} = dz + y$, given that $y(0) = 1, z(0) = -1$ for $y(0.1), z(0.1)$.

20.31 Describe Runge Kutta method (4th order) for obtaining solution of initial value problem:

$$y^{"} = f(x.y.y'),\ y'(x_0) = y_0,\ y'(x_0) = y'_0$$

20.32 Obtain the value of $x(0.1)$, given

$$\frac{d^2y}{dx^2} = t\frac{dx}{dt} - 4x,\ \ x'(0) = 0$$

20.33 Solve $dy/dx = -xz$, $dz/dx = y^2$,given $y(0) = 1$, $z(0) = 1$, for $x = 0, (0.2)(0.4)$.

20.34 Solve $y" - x(y')^2 + y^2 = 0$ using Runge Kutta method for $x = 0.2$ given $y(0) = 1, y'(0) = 0$, taking $h = 0.2$.

20.35 Computer the value of $y(0.2)$ given $y" = -y, y(0) = 1, y'(0) = 0$.

20.36 Find $y(0.1), y(0.2), z(0.2)$ given the differential equations $\frac{dy}{dx} = x + z, \frac{dz}{dx} = x - y^2$. The intitial values are $y = 2, z = 1$ when $x = 0$.

20.37 Evalute $x(0.1), y(0.1), x(0.2), y(0.2)$ given the differential equations $\frac{dx}{dt} = ty + 1, \frac{dy}{dt} = -tx$, the initial value are $x = 0, y = 1$ when $t = 0$.

20.38 Find $y(0.3)$,given the differential equations $\frac{dy}{dx} = 1 + xz$, $\frac{dz}{dx} = -xy^2$, the intial value are $y = 2$, $z = 1$ when $x = 0$.

20.39 Find $y(0.1)$ given $y" = y^3, y(0) = 10, y'(0) = 5$ by Runge Kutta method .

20.40 Find $y(0.1)$ given $y" + 2xy' - 4y = 0, y(0) = 0.2, y'(0) = 0.5$.

APPENDIX A

EXACT DIFFERENTIAL EQUATIONS

Exact Differential Equations

A first order differential equation of the form

$$Mdx + Ndy = 0$$

where M and N are functions of x, y, is said to be exact, if left hand side is total differential (i.e., $du = 0$). To be exact (total) differential M and N must hold

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x} = \frac{\partial^2 u}{\partial y \partial x}$$

Solving Method

The function $u(x, y)$ can be found by hit and trial procedure or in following way. Differentiate $Mdx + Ndy$ with respect to x treating y as constant

$$u = \int Mdx + f(y)$$

where $f(y)$ is constant of integration and is function of y only. To determine $f(y)$ we derive $\dfrac{\partial u}{\partial y}$ and compare with $N = \dfrac{\partial u}{\partial y}$.

■ EXAMPLE 1.1

Solve $2xdx + 2ydy = 0$

Solution: Here $M = x$, therefore

$$u = \int 2xdx + f(y) = x^2 + f(y)$$

Now, $\dfrac{\partial u}{\partial y} = f'(y)$ and $N = 2y$, this gives $f'(y) = 2y$, which gives $f(y) = \int 2ydy = y^2 + c$. Therefore $u = x^2 + y^2 + c$. □

Shortcut Method

Compute I part by intergrate M with respect to x treating y as constant and then II part by integrating only those terms of N which are free from x. Finally, $u = I + II + c$

■ EXAMPLE 1.2

Solve $2xdx + 2ydy = 0$

Solution:

$$I = \int 2xdx = x^2$$

$$II = \int 2ydy = y^2$$

$$u = I + II + c = x^2 + y^2 + c$$

□

APPENDIX B

ANSWERS OF SELECTED PROBLEMS

1.1.a Here, $u = x^2$ and $v = y^2$, this implies

$$u_x = 2x, \; u_y = 0 \; v_x = 0 \; v_y = 2y$$

Cauchy Riemann equations give,

$$u_x = v_y \quad \Rightarrow 2x = 2y \; \text{ and } u_y = 0 = -v_x$$

Function is nowhere analytic except at $x = y$.

1.1.b Check Cauchy Riemann equations. Not satisfied. Not analytic.

1.1.d Aanalytic everywhere except $z = \pm 1$.

1.2 Here,

$$u = 4x + y \; \Rightarrow u_x = 4, \; u_y = 1$$

and

$$v = -x + 4y \; \Rightarrow v_x = -1, \; v_y = 4$$

Cauchy Riemann equations are satisfied. Partial derivatives u_x, u_y, v_x, v_y exist and are continuous, therefore given function is differentiable everywhere. We have

$$\frac{df}{dz} = \frac{\partial f}{\partial z} = u_x + iv_x = 4 - i$$

1.3 Here, $w = \rho(\cos\phi + i\sin\phi) = \rho e^{i\phi}$ and

$$z = \ln\rho + i\phi = \ln\rho + \ln(e^{i\phi}) = \ln\rho e^{i\phi} \Rightarrow e^z = \rho e^{i\phi} = w \Rightarrow z = \ln w$$

Now problem reduces to find the value of z for which function $w = e^z$ is not analytic.

Important: To find z for which function ceases to be analytic, solve $\frac{dz}{dw} = 0$, i.e.,

$$\frac{dz}{dw} = \frac{1}{w} = e^{-z}$$

Thus $\frac{dz}{dw} = 0 \Rightarrow e^{-z} = 0$ which gives $z = \infty$

1.5 for $z = 0$ function ceases to be analytic.

1.8 Here, $f(z) = z^3 = (x + iy^3)$, which gives

$$u = x^3 - 3xy^2 \text{ and } v = 3x^2y - y^3$$

Compute Cauchy Riemann equations and check these holds. Hence function is analytic.

1.9 Here, $w = \sin z = \sin(x+iy) = \sin x \cosh y + i\cos x \sinh y$ (Since $\cos(iy) = \cosh y$ and $\sin(iy) = i\sinh y$).

$$u_x = \cos x \cosh y \quad u_y = \sin x \sinh y$$
$$v_x = -\sin x \sinh y \quad v_y = \cos x \cosh y$$

Cauchy Riemann equations are satisfied. u_x, u_y, v_x, v_y exists and are continuous. Hence w is analytic function. Now,

$$\frac{dw}{dz} = u_x + iv_x = \cos x \cosh y - \sin x \sinh y = \cos(x+iy) = \cos z$$

1.11 Here,

$$u_x = 2x - 2y, \quad u_y = 2ay - 2x \quad v_x = 2bx + 2y \quad v_y = -2y + 2x$$

Now from Cauchy Riemann equations, we get

$$u_x = v_y \quad \Rightarrow 2x - 2y = -2y + 2x$$

which is true, and

$$u_y = -v_x \quad \Rightarrow 2ay - 2x = -(2bx + 2y)$$

Compare coefficients of x and y on both side

$$a = -1, \ b = 1$$

Now,

$$f'(z) = u_x + iv_x = (2x - 2y) + i(2x + 2y) = 2(1+i)z$$

1.12 Here $f(z) = z|z| = (x+iy)|x+iy| = x\sqrt{(x^2+y^2)} + iy\sqrt{(x^2+y^2)}$, i.e.,

$$u = x\sqrt{(x^2+y^2)}, \quad v = y\sqrt{(x^2+y^2)}$$

Compute u_x, u_y, v_x, v_y and check, $u_x \neq v_y$ and $u_y \neq -v_x$, therefore function is nowhere analytic.

1.15

$$u_x = u_y = v_x = v_y = 0$$

It is obvious that the Cauchy Riemann equations are satisfied at $z = 0$, i.e., at $x = 0$, $y = 0$. But derivative along $y = mx$ at $z = 0$ is

$$f'(0) = \lim_{x} \to 0 \frac{f(z) - f(0)}{z} = \lim_{x} \to 0 \frac{\sqrt{xy} - 0}{x + iy} = \lim_{x} \to 0 \frac{\sqrt{x.mx} - 0}{x + imx} = \lim_{x} \to 0 \frac{\sqrt{m}}{1 + im}$$

Evidently, this limit depends on m, which differers for different values of m. i.e., $f'(0)$ is not unique. This shows $f'(0)$ does not exist. Hence given function is not analytic.

1.16.a $v = x^2 - y^2 + 2y$

1.16.b $2y - 3x^2y + y^3$

1.16.c

$$u_x = \frac{x}{x^2+y^2} \text{ and } u_y = \frac{y}{x^2+y^2}$$
$$u_{xx} = \frac{y^2-x^2}{(x^2+y^2)^2} \text{ and } u_{yy} = \frac{x^2-y^2}{(x^2+y^2)^2}$$
$$u_{xx}+u_{yy} = 0, \text{ Therefore function is harmonic.}$$

Now,

$$dv = v_x dx + v_y dy = -u_y dx + u_x dy = \frac{xdy - ydx}{x^2+y^2}$$
$$v = tan^{-1}\frac{y}{x}$$

1.16.d $v = 3x^2y + 6xy - y^3$

1.17.a $w = z^2 + (5-i)z - \frac{i}{z}$

1.17.b $w = \cos z$

1.17.c $w = 2z^2 - iz^3$

1.17.d $w = ize^{-z}$

1.17.e $w = ze^{2z}$

1.17.f $w = 2i\log z - (2-i)z$

1.17.g $w = \sin(iz)$

1.17.h $w = (1+i)\frac{1}{z}$

1.17.i $w = z + \frac{1}{z}$

1.18 Let $U = u - v$ and $V = u + v$, therefore $F(z) = U + iV = (1+i)(u+iv) = (1+i)f(z)$. Now

$$U = \frac{e^y - \cos x + \sin x}{\cosh y - \cos x} = 1 + \frac{\sin x + \sinh y}{\cosh y - \cos x} \quad (\because e^y = \cosh y + \sinh y)$$

By Milne's method,

$$F(z) = \int [\phi_1(z,0) - i\phi_2(z,0)] + C$$
$$= (1+i)\int \frac{dz}{1-\cos z} = \frac{(1+i)}{2}\int \text{cosec}^2\frac{z}{2}dz + C$$
$$= (1+i)\cot\frac{z}{2} + C$$

Therefore,

$$f(z) = \cot\frac{z}{2} + C$$

To evaluate C, use condition $f\left(\frac{\pi}{2}\right) = \frac{3-i}{2}$, we get

$$C = \frac{1-i}{2}$$

Hence

$$f(z) = \cot\frac{z}{2} + \frac{1-i}{2}$$

1.19 Here, $v = r^2\cos 2\theta - r\cos\theta + 2$,

$$u_r = \frac{1}{r}v_\theta = -2r^2\sin 2\theta + r\sin\theta$$

and

$$u_\theta = rv_r = r(2r\cos 2\theta - \cos\theta)$$

$$\begin{aligned} du &= u_r dr + u_\theta d\theta \\ &= (-2r^2\sin 2\theta + r\sin\theta)dr + (2r^2\cos 2\theta - r\cos\theta)d\theta \end{aligned}$$

$$\Rightarrow u = -r^2\sin 2\theta + r\sin\theta + C$$

Now, $f(z) = u + iv = -r^2\sin 2\theta + r\sin\theta + C + i(r^2\cos 2\theta - r\cos\theta + 2)$. On arranging, we get

$$f(z) = i(z^2 - z) + 2i + C$$

1.20 $f(z) = (1+i)z^2 + (-2+i)z - 1$

1.21 $f(z) = \log iz$

1.22 Let $f(z) = u + iv$, so that $|f(z)|^2 = u^2 + v^2 = \phi(x,y)$ (say)

$$\phi_x = 2uu_x + 2vv_x, \quad \phi_y = 2uu_y + 2vv_y$$

and

$$\phi_{xx} = 2\left[uu_{xx} + u_x^2 + vv_{xx} + v_x^2\right], \quad \phi_{yy} = 2\left[uu_{yy} + u_y^2 + vv_{yy} + v_y^2\right]$$

This gives

$$\phi_{xx} + \phi_{yy} = 2\left[u(u_{xx} + u_{yy}) + (u_x^2 + u_y^2) + v(v_{xx} + v_{yy}) + (v_x^2 + v_y^2)\right]$$

Since CR equations are satisfied here and Laplace equation also holds for u and v, therefore

$$\phi_{xx} + \phi_{yy} = 4\left[(u_x^2 + v_x^2)\right] = 4|f'(z)|^2$$

2.1.a

$$\frac{1}{3}(i-1)$$

2.1.b

$$-\frac{1}{2} + \frac{5}{6}i$$

2.2.a Equation of line AB

$$y = 7x - 6$$

Now

$$\int f(z)dz = \int (x^2 + ixy)(dx + idy)$$

substitute $y = 7x - 6$, $dy = 7dx$, then on integrating

$$= \frac{1}{3}[-147 + 71i]$$

2.2.b Here

$$\int f(z)dz = \int (x^2 + ixy)(dx + idy)$$

substitute $x = t, y = t^3$, $dx = dt$ and $dy = 3t^2dt$, then on integrating with respect to t from 1 to 2

$$= -\frac{1094}{21} + \frac{124}{5}i$$

2.3 -1

2.4.a (i) 0 (ii) $2\pi i$

2.4.b $2\pi ie$

2.4.c $-2\pi i$

2.4.d $4\pi i$

2.4.e $\frac{i\pi e^2}{12}$

2.4.f $\frac{8i\pi e^{-2}}{3}$

2.4.g $4\pi i$

2.5 $-12\pi i$

2.6 $-\frac{4i\pi}{9}$

2.7 $3\pi i$

3.1 Using partial fraction, rewrite the function as

$$\begin{aligned} f(z) &= -\frac{1}{2}\left(1 - \frac{z}{2}\right)^{-1} - \frac{1}{z}\left(1 - \frac{1}{z}\right)^{-1} \\ &= -\frac{1}{2} - \frac{z}{4} - \frac{z^2}{8} - \ldots - \frac{1}{z} - \frac{1}{z^2} - \frac{1}{z^3} \end{aligned}$$

3.2 Using partial fraction, rewrite the function as

$$\begin{aligned} f(z) &= \frac{1}{5}\left(\frac{z-2}{1+z^2}\right) + \frac{1}{5}\left(\frac{1}{z+2}\right) \\ &= \frac{z-2}{5z^2}\left(1 + \frac{1}{z^2}\right)^{-1} + \frac{1}{10}\left(1 + \frac{z}{2}\right)^{-1} \\ &= \frac{1}{5}\left[\ldots - 2z^{-8} + z^{-7} + 2z^{-6} - z^{-5} - 2z^{-4} + z^{-3} + 2z^{-2} - z^{-1}\right] \\ &\quad + \frac{1}{5}\left[\frac{1}{2} - \frac{z}{4} + \frac{z^2}{8} - \frac{z^3}{16} + \ldots\right] \end{aligned}$$

3.3 Put $z - 1 = t$. Then expand series in powers of t finally use back substitution to get

$$e\left[\frac{1}{(z-1)^2} + \frac{1}{(z-1)} + \frac{1}{2!} + \frac{z-1}{3!} + \cdots\right]$$

3.4

$$b_n = a_n = \frac{1}{2\pi}\int_0^{2\pi} \sin(2c\cos\theta)\cos n\theta d\theta$$

and

$$f(z) = a_0 + \sum_{n=1}^{\infty} a_n (z^n + z^{-n})$$

3.5 $f(z)$ is analytic at $z = 0$, therefore by Taylor's theorem

$$a_n = \frac{1}{2\pi} \int_0^{2\pi} \cos(c \sin\theta - n\theta) d\theta$$

and

$$b_n = (-1)^n a_n$$

therefore

$$f(z) = \sum_{-\infty} \infty a_n z^n$$

3.6 (a) $-1 + 2(z - z^2 + z^3 - z^4 + \ldots)$ (b) $\frac{1}{2}(z-1) - \frac{1}{2^2}(z-1)^2 + \frac{1}{2^3}(z-1)^3$

3.7 (a) $-\dfrac{1}{2z^4} + \dfrac{1}{2z^3} - \dfrac{1}{2z^2} + \dfrac{1}{2z} - \dfrac{1}{6} + \dfrac{z}{18} - \dfrac{z^2}{54} + \dfrac{z^3}{162} \ldots\ldots$

(b) $\dfrac{1}{z^2} - \dfrac{4}{z^3} + \dfrac{13}{z^4} - \dfrac{50}{z^5} + \ldots$

(c) $\dfrac{1}{2(1+z)} - \dfrac{1}{4} + \dfrac{(z+1)}{8} - \dfrac{(z+1)^2}{16} + \ldots.$

(d) $\dfrac{1}{3} - \dfrac{4}{9}z + \dfrac{13}{27}z^2 - \dfrac{40}{81}z^3 + \ldots$

3.9 $\frac{1}{10}[(\frac{2}{z} + \frac{2}{z^2} + \frac{2}{z^5} + \ldots) - (\frac{z}{2} + \frac{z^3}{8} + \ldots)]$

3.10 (i) $-\frac{1}{2z} + \sum_{n=0}^{\infty} \left[(-1)^{n+1} \frac{1}{2^{n+2}} - \frac{1}{2^{n+1}}\right] z^n$

(ii) $-\frac{1}{2z} - \frac{1}{2z} \sum_{n=0}^{\infty} (-1)^n (\frac{2}{z})^n - \frac{1}{3} \sum_{n=0}^{\infty} (\frac{z}{3})^n$

3.11 when $|z-1| > 4$, $\frac{1}{z-1} - \frac{2}{(z-1)^2} + \frac{9}{(z-1)^3} + \frac{36}{(z-1)^4} + \ldots$

when $0 < |z-1| < 4$, $\frac{1}{4}[1 - \frac{1}{(z-1)^2} + \frac{7}{4}\frac{1}{z-1} + \frac{9}{16} - \frac{9}{-64}(z-1) \ldots]$

3.13 $\frac{2}{z+2} + \frac{3}{(z+2)^2} + \frac{3^2}{(z+2)^3} + \cdots + \frac{1}{5}\left[1 + \frac{z+2}{5} + \frac{(z+2)^2}{5^2} + \frac{(z+2)^3}{5^3} + \cdots\right]$

4.1 For zeros

$$\frac{z-2}{z^2} \sin\frac{1}{z-1} = 0$$

$$z - 2 = 0 \text{ and } \sin\frac{1}{z-1} = 0$$

Therefore $z = 2$ and $\frac{1}{z-1} = n\pi$, i.e.

$$z = 2 \text{ and } z = \frac{1}{n\pi} + 1$$

It is obvious that $z = 2$ is simple zero. For $z = \frac{1}{n\pi} + 1$, limit point of zero when $n \to \infty$ is 1, i.e., $z = 1$ is isolated essential singularity. Except these $z^2 = 0$ provides pole at $z = 0$ of order 2.

4.2.b Poles: $\sin z - \cos z = 0 \Rightarrow z = \frac{\pi}{4}$

4.2.d Here $f(z) = \dfrac{\cot\pi z}{(z-a)^2} = \dfrac{\cos\pi z}{\sin\pi z (z-a)^2}$, The poles are given by putting denominator equal to zero. i.e.,

$$\sin\pi z (z-a)^2 = 0$$

$$(z-a)^2 = 0 \Rightarrow z = a \text{ Double pole}$$

and

$$\sin \pi z = 0 \Rightarrow \pi z = n\pi \Rightarrow z = n$$

Limit point of pole $z = n$, when $n \to \infty$ is ∞. Thus $z = \infty$ is non-isolated essential singularity.

4.2.j Non-isolated essential singularity.

5.1.a $\frac{\pi}{2\sqrt{5}}$

5.1.b 0 ; **Hint**

$$I = \int_0^{2\pi} \frac{\cos\theta}{3+\sin\theta} d\theta = Real part of \int_0^{2\pi} \frac{e^{i\theta}}{3+\sin\theta} d\theta$$

5.1.c 0

5.1.d $\frac{2\pi}{\sqrt{3}}$

5.1.e $\frac{\pi}{12}$

5.1.f Let

$$I = \int_0^{2\pi} e^{\cos\theta}[\cos(\sin\theta - n\theta + i\sin(\sin\theta - n\theta)]d\theta$$
$$= \int_0^{2\pi} e^{e^{i\theta}} e^{-in\theta}$$

Put $z = e^{i\theta}$,

$$I = \int_C \frac{e^z}{iz^{n+1}} dz$$

Now by residue theorem,

$$I = \frac{2\pi}{n!}$$

Now compare real parts to obtain

$$\int_0^{2\pi} e^{\cos\theta}[\cos(\sin\theta - n\theta)]d\theta = \frac{2\pi}{n!}$$

5.2.a $\frac{\pi e^{-m}}{2}$

5.2.b $\pi \log 2$

5.2.c $\frac{3\pi}{16}$

5.2.d $\frac{\pi}{3}$

6.1 $\bar{x} = 20$, $\mu_1 = 0$, $\mu_2 = 21.67$, $\mu_3 = 30$, $\mu_4 = 765.67$

6.2 $\mu_1' = 0$, $\mu_2' = 109.09$, $\mu_3' = 0$

6.3 $\mu_1 = 0$, $\mu_2 = 33.7$, $\mu_3 = -102.7$, $\mu_4 = 2567.5$

6.4 $\sqrt{\beta_1} = 0.074, \beta_2 = 2.03$

6.5 $\beta_2 = 2.408$ (Platykurtic)

6.6 $\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 0$, $\mu_4 = 11$, $\beta_1 = 0$, $\beta_2 = 2.75$

6.7 $\mu_1' = 5$, $\mu_2' = 31$, $\mu_3' = 201$

6.8 $\bar{x} = \mu_1' = 5$, $\mu_1 = 0$ $\mu_2 = 3$, $\mu_3 = 0$, $\mu_4 = 26$, $\beta_1 = 0$, $\beta_2 = 2.89$

6.9 $\mu_1 = 0$ $\mu_2 = 1.5$, $\mu_3 = 0$, $\mu_4 = 6$, $\nu_1 = 3$, $\nu_2 = 10.5$, $\nu_3 = 40.5$, $\nu_4 = 168$

6.10 $\beta_1 = 2.25$, $\beta_2 = 2.75$

6.11 $\mu_1 = 0$ $\mu_2 = 16$, $\mu_3 = -64$, $\mu_4 = 262$, $\beta_1 = 1$, $beta_2 = 1.02$

6.12 Bowley's Coefficient of skewness=-0.16, $\gamma_1 = -0.15$

6.13 $\bar{x} = 19.32$, $Z = 20.62$, $\sigma = 8.22$, $S_k = -0.16$

6.14 $\bar{x} = 715.5$, $Z = 669.23$, $\sigma = 190.15$, $S_k = 0.243$

6.15 $\bar{x} = 58.35$, $Z = 59.17$, $\sigma = 6.98$, $S_k = 0.12$

6.16 -0.39

6.17 $Q_1 = 12, Q_2 = 12, Q_3 = 13, S_{k_b} = 1$

6.18 $Q_1 = 10.42, Q_2 = 20, Q_3 = 27.81, S_{k_b} = -0.102$

6.19 $Q_1 = 30, Q_2 = 38.44, Q_3 = 46, S_{k_b} = -0.08$

6.20 $Q_1 = 37.4, Q_2 = 50.3, Q_3 = 60.8, S_{k_b} = -0.103$

6.21 For A $S_{k_b} = -0.013$, For B -0.059; B is more skewd.

7.1 $y = 2.022x + 0.503$

7.2 $y = x + 1.9$

7.3 $y = 0.0476x + 0.0041$

7.4 $y = 54.4917x + 0.5160$

7.5 $y = 45.7267x + 6.1624$

7.6 $y = 8.6429x - 1.6071$

7.7 $y = -11.8005x + 5.3041$

7.9 $y = 1.48 + 1.13(x-2) + 0.55(x-2)^2$

7.12 $y = e^{0.5x}$

7.13 $V = 1.46.3e^{-0.412t}$

7.15 $y = 7.17x^{1.95}$

7.19 $y = 12.6751 + \dfrac{8.2676}{x} - \dfrac{5.7071}{x^2}$

7.20 $y = 2.396x + \frac{3.029}{x}$

7.21 $xy = 21.3810x + 15.7441$

7.22 $y = 1.973x - 0.2984x^2$

7.23 Treat γ as independent variable. $pv^{1.276} = 1.039$

8.1.a $\bar{(X)} = 43,\ \bar{Y} = 59,\ r = 0.12$

8.1.b $\bar{(X)} = 40,\ \bar{Y} = 34,\ r = 0.99$

8.2 $r = 0.79$

8.3 $r = 0.995$

8.4 $r = 0.997$

8.5 $r = 1$

8.6 $r_{xy} = 0.96,\ r_{uv} = -0.96$

8.7 $r = 0.61$

8.8 $r = 0$

8.9 $r = -0.954$

8.10 $r = 0.787$

8.11 $r = 0.947$

8.12 $r = -0.93$

8.13 $r = -0.774$

8.14 $r = -0.99$

8.15 $r = 0.24$

8.16 $r = 0.82$

8.17 $r = -0.94$

8.18 Missing values: We have

$$\bar{X} = \frac{6+2+10+?+8}{5} = 6 \Rightarrow ? = 4$$

Missing in X=4, in Y =5 and $r = -0.92$

8.19 Missing in X=8, in Y =12 and $r = 0.96$

8.20 $r = 0.8$

8.21 $r = -0.32$

8.22 $r_xy = 0.3$, No change

8.23 $r = -0.9$

8.24 $r = 0.52$

8.25 $r = 0.42$

8.26 $r = 0.81$

8.27 $r = 0.76$, High degree positive correlation.

8.28 $r = 0.12$ liking of judges are similar with low degree positive correlation.

8.29 $r = 0.51$

8.30 $r = -0.905$

8.31 $r = 0.93$

8.32 $r = 0.918$

8.33 $r = 0.733$

8.34 $r = 0.643$

8.35 $r = 0.139$

8.36 $r = -0.93$

8.37 $r = 0.73$

8.38 $r = -0.11$

9.1 16.30

9.3 $x = 0.3675y + 22.36$, $y = 0.653x + 21.88$; Economics marks, $y = 55$

9.4 $x = 1.6y - 60$, $y = 0.4x + 51$ (a) 20 (b) 83

9.5 $x = 0.268y + 26.743$, $y = 0.657x + 21.648$, x=42, y=55

9.7 $y = -1 + 0.75x$, Salaries = Rs 800.

9.10 $r = 0.904$, $x = 2.37y - 74.83$, $y = 0.345x - 36.575$

9.11 $x = -0.667y + 1.3$, $y = 0.611x + 9.3$

9.13 $\sigma_x = 3$, $r_{xy} = 0.75$

9.14 $\sigma_x = \frac{3}{4}\sigma_y$, $b_{xy} = 0.6$, $r = 0.8$, $\overline{(x)} = 62.4$

9.15 (i) $\bar{x} = 1, \bar{y} = 2$,
(ii) $\sigma_y^2 = 4$,
(iii) $r_{xy} = 0.86$

9.16 First line is line of regression of y on x.

9.19 (i) $\bar{x} = 6$, (ii) $\sigma_y = 13.3$ (iii) $b_{yx} = \frac{64}{45}$ (iv) $r = 0.53$

9.20 (i) $b_{xy} = -0.4$ (ii) $y = 34.5$

9.21 Hint: Student may take help from chapter of curve fitting. Just fit a straight line for given data. $y = 8.8 + 1.015x$ and y at $x = 16$ = 25.04

9.22 $r_{12.3} = 0.04$ and $r_{13.2} = 0.446$

9.23 $x_3 = 0.4x_1 + 0.424x_2$

9.24 $x_3 = 11.48 + 0.688x_1 + 0.6x_2$, $x_3 = 60.44$

9.25 $x_3 = 3.49 + 0.053x_1 + 0.963x_2$, $x_3 = 57.122$

9.26 $x_1 = 16.48 + 0.39x_2 - 0.62x_3$

10.9 $\frac{4}{5}$

10.10 Each (4) person can stay in 5 ways. The total number of ways = 5^4. They stay in different hotels in $5 \times 4 \times 3 \times 2$ ways. Hence Probability of staying in different hotels = $\frac{5\times4\times3\times2}{5^4} = \frac{24}{125}$

10.11 $\frac{^5C_2}{^8C_2} + \frac{^3C_2}{^8C_2} = \frac{13}{28}$

10.12 (i) $\frac{^3C_1 \times ^7C_1 \times ^3C_2}{^10C_2} = \frac{8}{15} (ii) \frac{^3C_2}{^10C_2} = \frac{1}{15}$

10.13 With replacement (i) $\frac{15}{32}$ (ii) $\frac{17}{32}$ Without replacement (i) $\frac{15}{28}$ (ii) $\frac{13}{28}$

10.14 (i) Total balls = 7+3=10
The number of ways of drawing 3 balls $= 10^3$
The number of ways drawing 3 white balls $= 7^3$
$\therefore$ P(Drawing three white balls)$= \frac{7^3}{10^3} = \frac{343}{1000}$
and P(at least one black ball)$= 1 - P$ (all are white)

$$= 1 - \frac{343}{1000} = \frac{657}{1000}$$

(ii) The first ball and the last ball are of different colors. Following cases are possible
WWB, WBB, BBW, BWW
P(The first ball and the last ball are of different colors)

$$\begin{aligned} &= \frac{7}{10} \times \frac{7}{10} \times \frac{3}{10} + \frac{7}{10} \times \frac{3}{10} \times \frac{3}{10} + \frac{3}{10} \times \frac{3}{10} \times \frac{7}{10} + \frac{3}{10} \times \frac{7}{10} \times \frac{7}{10} \\ &= \frac{21}{50} \end{aligned}$$

10.15 $\frac{^7C_2}{^{12}C_2}$

10.16 $\frac{1}{2} \times \frac{^6C_1 \times ^4C_1}{^{10}C_2} + \frac{1}{2} \times \frac{^4C_1 \times ^8C_1}{^{12}C_2} = 0.51$

10.17 $\frac{^4C_3}{^9C_3}$

10.18 $\frac{^4C_1 \times ^5C_1 \times ^6C_1}{^{52}C_3} = \frac{24}{91}$

10.19

$$\begin{aligned} p(B_1/A) &= \frac{P(B_1)P(A/B_1)}{P(B_1/A)P(A/B_1) + P(B_2)P(A/B_2)} \\ &= \frac{1}{2} \end{aligned}$$

10.20 Total number of insured drivers =2,000+4,000+6,000=12,000 Let B_1: scooter driver, B_2: Car driver, B_3: Truck driver $P(B_1) = \frac{2,000}{12,000} = \frac{1}{6}$, $P(B_2) = \frac{4,000}{12,000} = \frac{1}{3}$, $P(B_3) = \frac{6,000}{12,000} = \frac{1}{2}$

Required probability

$$P(B_2/A) = \frac{p(B2)p(A/B2)}{p(B1)p(A/B1) + p(B2)p(A/B2) + p(B3)p(A/B3)}$$
$$= \frac{\frac{1}{3} \times 0.03}{\frac{1}{6}0.01 + \frac{1}{3} \times 0.03 + \frac{1}{2} \times 0.015} = \frac{3}{26}$$

10.21 Let B_1=a student is well prepard, B_2=a student is unprepared, S=a student is pass, F=a student is fail

$p(B1) = 90\% = \frac{90}{100}$, $p(B2) = 10\% = \frac{10}{100}$, $p(S/B1) = 0.85 = \frac{85}{100}$
$p(F/B1) = 1 - \frac{85}{100} = \frac{15}{100}$ $p(F/B2) = 0.60 = \frac{60}{100}$ $p(S/B2) = 1 - \frac{60}{100} = \frac{40}{100}$
According to bayes' Theorem
(i)

$$p(B1/S) = \frac{p(B1)p(S/B1)}{p(B1)p(S/B1) + p(B2)p(S/B2)}$$
$$= \frac{\frac{90}{100} \times \frac{85}{100}}{\frac{90}{100} \times \frac{85}{100} + \frac{10}{1001} \times \frac{40}{100}} == 0.95$$

(ii)

$$p(B2/F) = \frac{p(B2)p(F/B2)}{p(B1)p(F/B1) + p(B2)p(F/B2)}$$
$$= \frac{\frac{10}{100} \times \frac{60}{100}}{\frac{90}{100} \times \frac{15}{100} + \frac{10}{100} \times \frac{60}{100}} = 0.31$$

10.22 P(the selected fountain pen is manufactured by machine A),$= \frac{30}{100}$
P(The selected fountain pen is manufactured by machine B),$= \frac{30}{100}$
P(The selected fountain pen is manufactured by machine C),$= \frac{40}{100}$
$P(A/D)$=The produce of A is defective $= \frac{4}{100}$
$P(B/D)$=The selected produce of B is defective $= \frac{5}{100}$
$P(C/D)$=The produce of C is defective $= \frac{10}{100}$
According to Bayes' Theorem ,

$$P(D/C) = \frac{\frac{40}{100} \times \frac{10}{100}}{\frac{30}{100} \times \frac{4}{100} + \frac{30}{100} \times \frac{5}{100} + \frac{40}{100} \times \frac{10}{100}} = \frac{40}{67}$$

12.1 $\sigma = 0.024$, Limits 0.353 ¡ p ¡ 0.447, Coin is symmetrical.

12.2 $\sigma = 0.012$, Limits $60.4\% - 67.6\%$

12.3 Yes, fair.

12.4 $\sigma_{p_1 \sim p_2} = 0.0236$, Yes these data reveal a significant difference between town A and town B, so far as the proportion of wheat consumers is concerned.

12.7 $z = 3.26$, Significant.

12.8 48.75-51.25

12.9 No. Since Standard Error = 0.097

12.10 $z = 10.87$, Significant

12.11 74.625-75.375

12.12 $\sigma_{\bar{x_1}-\bar{x_2}} = 1.419$. Inconsistent.

12.13 $t = 1.89$, Not significant. Mean may be 66 inch.

12.14 Assuming independent samples, $t = 0.79$

12.15 $t = 3.333$, difference significant.

12.16 Soldiers are on average taller than sailors.

12.17 $t = 1.48$, No benefit.

12.19 Not taken from same population.

12.21 (i) $t = 0.827$ (ii) $t = 0.612$; Not significant in both cases.

12.22 $\chi^2 = 10.6, \chi^2_{0.05}(1) = 3.841$, Quinine prevents malaria.

12.23 $\chi^2 = 7.476$, Credit squeeze is effective.

12.24 $\chi^2 = 10.3$, B is not superior than A.

12.25 $\chi^2 = 23.67$, Results do not commensurate with the general examination results.

13.1 $F_c = 1.17 < 7.71$, Not significant.

13.2 $F_c = 3.01 < F_{0.05} = 8.74$, Not significant.

13.5 $F_c = 9 > F_{0.05} = 5.14$, Significant.

13.6 Not significant.

13.7 $F_c = 1.125 < F_{0.05}$, Not significant.

13.9 F_c between Salesman = 3.58 and F_c Between territory=2.39, both are less than table value, hence difference in both is not significant. Not significant.

13.10 Difference is significant with regard to yield (Here $F = 8.14 > F_{0.5} = 3.88$)

13.11 (a) Not significant. (b) Significant.

15.1 CL = 26.5, UCL = 34.3, LCL = 18.7

15.2 CL=2.5, UCL=2.503, LCL=2.497; Process is out of control.

15.3 CL=0.5, UCL=0.53, LCL=0.47

15.4 CL=99.6, UCL=1.03.65, LCL=95.55; Process is out of control.

15.5 Mean of sample mean = 73.6, Mean of range = 62.2 UCL=109.7, LCL=37.5

15.7 CL=6, UCL=14.33, LCL=0.

15.8 $\overline{\overline{X}} = 44.2$, $\overline{R} = 5.8$,
For $\overline{X}$-chart: UCL=47.56, LCL=40.84.
For R-chart: UCL=12.267, LCL=0.

15.9 For $\overline{X}$-chart: CL= 0.0279, UCL=0.041, LCL=0.015
For R-chart: CL= 0.0227, UCL=0.048, LCL=0.

15.11 $p = \frac{2}{100} = 0.02$, $n = 100, q = 1 - 0.02 = 0.98$
$\text{CL} = n.p = 100 \times 0.02 = 2$
$\text{UCL} = np + \sqrt{npq} = 2 + 3\sqrt{100 \times 0.02 \times 0.98} = 6.2$
$\text{LCL} = np - 3\sqrt{npq} = 2 - 4.2 = -2.2 = 0$

15.12 $\bar{p} = 0.11$, UCL=0.204, LCL=0.016

15.13 $\bar{p} = 2\%$, $UCL = 6.2\%$, $LCL = 0$

15.14 np-chart : CL=4, UCL=9.88, LCL=-1.88 (0)

15.15 np-chart: CL=10, UCL=19, LCL=1; Process is out of control

15.16 CL=$\bar{c} = 16$, UCL=28, LCL=4

15.17 CL=$\bar{c} = 4$, UCL=10, LCL=-2 (0)

15.18 CL=$\bar{c} = 4$, UCL=10, LCL=-2 (0)

16.2.a Interval:[1,2]; root: 1.839 (10 steps)

16.2.b Interval:[0,1]; root: 0.7885 (14 steps)

16.2.c Root lies in [0,1]. Required root upto 3 decimal places is 0.567.

16.2.c Interval:[0,1]; root: 0.5671 (13 steps)

16.2.d Interval:[0,1]; root: 0.5885 (14 steps)

16.2.e Interval:[0,1]; root: 0.6190 (13 steps)

16.2.f Interval:[0,1]; root: 0.3918 (13 steps)

16.3.a Interval:[1,2]; root: 1.796

16.3.b Interval:[1,2]; root: 1.302

16.3.c Interval:[1,2]; root: 1.465

16.3.d Interval:[0,1]; root: 0.754

16.4 Interval:[1,2]; root: 1.8880

16.5 root: 0.739

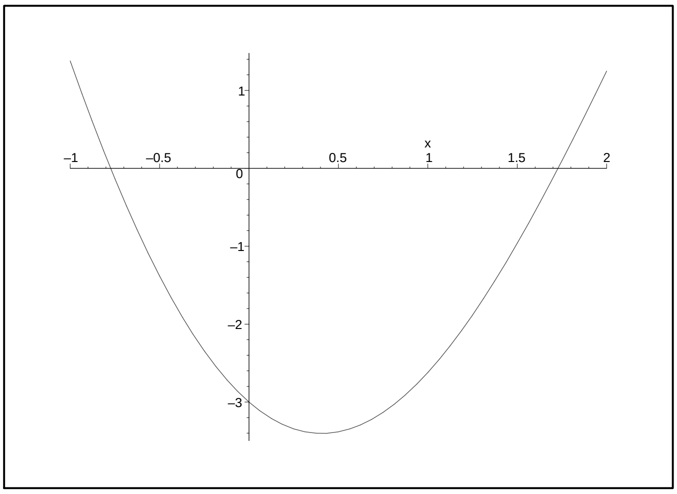
16.6 Positive root of 30, i.e. $\sqrt{30}$. Thus we have to find x such that

$$x = \sqrt{30}$$

$$\Rightarrow x^2 - 30 = 0$$

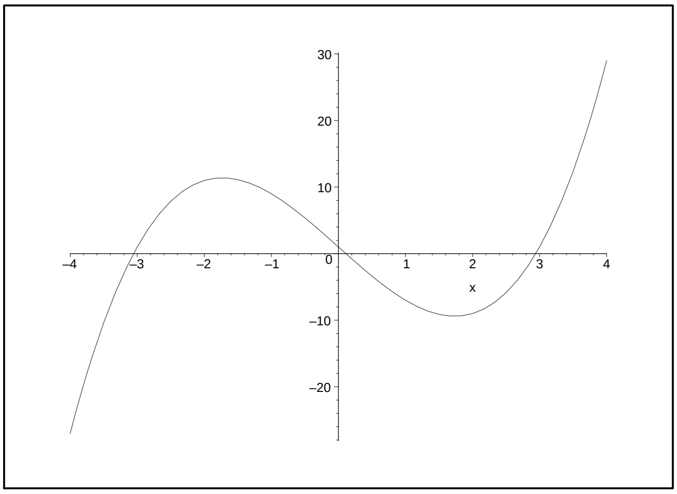
Now solve this equation using bisection method for *x*. In step 15, we get 5.47721

16.7 To understand the problem, we plot function (Students need not it to show in exams.)



There are two roots which lies in (-1,0) and (1,2), it is obvious from figure that the magnitude of root in (-1,0) is less than other. So we will apply bisection method in (-1,0). Required root is -0.7736 (Other root is 1.7280).

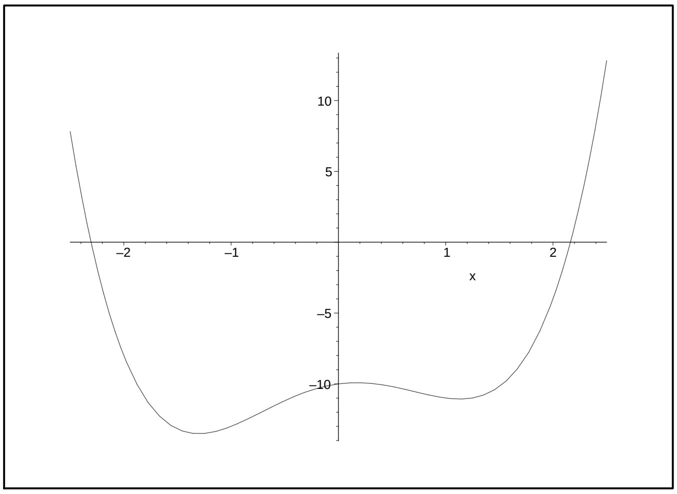
16.8 Plot function



Roots lies in (-4,-3)(0,1)(2,3). Smallest positive root lies in (0,1). That is 0.111 (in 12 steps)

16.9 2.02 (in 9 steps)

16.12.b Plot function



Decide yourself. approximated root after two iteration is 2.25.

16.13.a (2,3);2.0945

16.13.b (1,2);1.888

16.13.c (0,1); 1.9002

16.13.d (0,1);0.6566

16.14 2.5756

16.15 2.6207

16.16 4.6104

18.1 $(x,y,z) = (1,-2,4)$

18.2 $(x_1,x_2,x_3) = \left(\frac{-7}{4}, \frac{-19}{8}, \frac{21}{8}\right)$

18.3 $(x,y,z,w) = (2,2,2,1)$

18.4 $(x_1,x_2,x_3) = (6,-7,2)$

18.5 $(x,y,z) = \left(\frac{119}{4}, \frac{71}{4}, \frac{-79}{4}\right)$

18.6 $(x,y,z) = (5,-4,1)$

18.7 $(x,y,z) = (2.4254, 3.5730, 1.9259)$

18.8 $(x,y,z) = (3.000, 4.000, 5.000)$

18.9 $(x,y,z) = (0.9936, 1.5069, 1.8486)$

18.10 $(x,y,z) = (1.000, 1.000, 1.000)$

18.11 $(x,y,z) = (1.000, 1.000, 1.000)$

18.12 $(x,y,z) = (1.013, -1.996, 3.001)$

18.13 $(x,y,z) = (2.556, 1.722, -1.055)$

18.14 $(x,y,z) = (2, 0.9998, 2.9999)$

18.15 $(x,y,z) = (0.6666, 2.0000, 0.6666)$

18.16 $(x,y,z) = (0.83, 0.32, 1.07)$