Master of Business Administration

(Open and Distance Learning Mode)

Semester – I

Quantitative Methods

Centre for Distance and Online Education (CDOE)

DEVI AHILYA VISHWAVIDYALAYA, INDORE

"A+" Grade Accredited by NAAC

IET Campus, Khandwa Road, Indore - 452001

www.cdoedavv.ac.in

www.dde.dauniv.ac.in

Program Coordinator

Dr. Manishkant Arya
Centre for Distance and Online Education (CDOE)
Devi Ahilya Vishwavidyalaya, Indore – 452001

Content Design Committee

Dr. Sangita Jain
Institute of Management Studies
Devi Ahilya Vishwavidyalaya, Indore – 452001

Dr. Yamini Karmarkar
Institute of Management Studies
Devi Ahilya Vishwavidyalaya, Indore – 452001

Dr. Geeta Neema
International Institute of Professional Studies
Devi Ahilya Vishwavidyalaya, Indore – 452001

Dr. Manishkant Arya
Centre for Distance and Online Education (CDOE)
Devi Ahilya Vishwavidyalaya, Indore - 452001

Language Editors

Dr. Arti Sharan
Institute of Engineering & Technology
Devi Ahilya Vishwavidyalaya, Indore – 452001

Dr. Ruchi Singh
Institute of Engineering & Technology
Devi Ahilya Vishwavidyalaya, Indore – 452001

SLM Author(s)

Dr. Neelmegh Chaturvedi
MBA, PhD
EMRC, Devi Ahilya Vishwavidyalaya, Indore – 452001

Ms. Rashmi Modi
MBA
EMRC, Devi Ahilya Vishwavidyalaya, Indore – 452001

Edition : 2022 (Restricted Circulation)
Published by : Centre for Distance and Online Education (CDOE), Devi Ahilya Vishwavidyalaya
Printed at : University Press, Devi Ahilya Vishwavidyalaya, Indore – 452001

Quantitative Methods

SYLLABI-BOOK MAPPING TABLE

Quantitative Techniques

Syllabi	Mapping in Book
UNIT-I **Basic mathematics for management:** Functions—meaning, types and applications; arithmetic progression, geometric progression and matrices and their business applications.	**Unit 1:** Basic Mathematics for Management (**Pages: 3-76**)
UNIT-II Construction of frequency distributions and their analysis in the form of measures of central tendency and variations; types of measures, their relative merits, limitations and characteristics; skewness: meaning and coefficient of skewness.	**Unit 2:** Frequency Distribution and Skewness (**Pages: 77-152**)
UNIT-III **Correlation analysis:** Meaning and types of correlation, Karl Pearson's coefficient of correlation and Spearman's rank correlation; regression analysis—meaning and two lines of regression; relationship between correlation and regression coefficients; time series analysis—measurement of trend and seasonal variations; time series and forecasting.	**Unit 3:** Correlation and Regression Analyses (**Pages: 153-227**)
UNIT-IV **Probability:** Basic concepts and approaches, addition, multiplication and Bayes theorem; probability distribution—meaning, types and applications, binomial, Poisson, normal and exponential distributions.	**Unit 4:** Probability: Theory and Distribution (**Pages: 229-277**)

CONTENTS

INTRODUCTION

NOTES

Mathematics is the most precise language known to humans. The only way to make our decision-making process precise is to quantify the information we have and then perform calculations on it to arrive at desired solutions. Quantitative methods make all this possible for us. These methods help us assimilate information in a numerical format and perform operations on it to know the outcome, which in turn helps us in decision making.

Quantitative techniques or methods are of paramount importance in the business world, where our success entirely depends on our ability to take correct and timely decisions. Today, most of the business problems have found their quantitative representation, which makes this field of study more interesting and useful to the students and businesspersons alike. These methods are a great aid to the business strategists. In the present world, knowledge of these methods is a prerequisite to enter any business environment.

This book sets the foundation for you to learn and apply various quantitative techniques for solving various day-to-day problems of estimation and decision making. Unit 1 acquaints you with the basic mathematics for management. Unit 2 deals with frequency distribution and skewness. Unit 3 is all about correlation and regression analyses. Unit 4 teaches you all about probability. The book includes numerous examples and practice questions to help develop clear understanding of the concepts and techniques taught in these units.

The book follows the SIM format or the self-instructional mode wherein each Unit begins with an Introduction to the topic followed by an outline of the Unit Objectives. The detailed content is then presented in a simple and organized manner, interspersed with Check Your Progress questions to test the understanding of the students. A Summary along with a list of Key Terms and a set of Questions and Exercises is also provided at the end of each unit for effective recapitulation.

UNIT 1 BASIC MATHEMATICS FOR MANAGEMENT

NOTES

Structure

1.0 INTRODUCTION

In this unit, you will learn about basic mathematics for management, including functions, progressions and matrices. In mathematics, a function expresses the inherent idea that one quantity (input) completely determines another quantity (output). A function assigns a unique value to each input of a specified type. Progression in mathematics refers to arithmetic progression and geometric progression. The arithmetic progression, also termed as arithmetic sequence, refers

to a sequence of numbers in a manner in which the difference of any two successive members of the sequence is a constant. The geometric progression, or a geometric sequence, refers to a sequence of numbers in which each term is equal to the product of the preceding term and a constant number called common ratio. A matrix is a powerful mathematical tool. Its role in the solution of linear equations is widely accepted by various disciplines, including commerce and economics.

1.1 UNIT OBJECTIVES

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

- Define a function
- Explain the types and applications of functions
- Explain the meaning and applications of arithmetic progression
- Explain the meaning and applications of geometric progression
- Describe the applications of matrices in business

1.2 FUNCTIONS

In mathematics, a function expresses the inherent idea that one quantity (input) completely determines another quantity (output). A function or mapping from a set A to a set B is a 'method' which pairs elements of the set A with unique elements of the set B and we denote $f: A \to B$ to indicate that f is a function from the set A to the set B.

B is called the codomain of the function f and A is called its domain. Also, for each element a of A, f defines an element b of B. We write $a \xrightarrow{f} f(a)$ or $a \xrightarrow{f} b$, $a \in A$, $b \in B$.

Consider the following examples:

(i) The relation $f = \{(1, d), (2, c), (3, a)\}$ from $A = \{1, 2, 3\}$ to $B = \{a, c, d\}$ is a function from A to B. The domain of f is A and the codomain of f is B.

(ii) The relation $f = \{(a, b), (a, c), (b, d)\}$ from $A = \{a, b\}$ to $B = \{b, c, d\}$ is not a function.

Range of function: Let $f: A \to B$ be a function. The range of the function $R(f)=\{f(a)/a \in A\}$. (Note that $R(f) \leq B$.)

Notes:

1. From Example (i): $R(f)$ is $\{d, c, a\}$.
2. Let $f: IR^+ \to IR$ be $f(x) = x^2$ (R^+, the set of + ve real numbers). Clearly f is a function whose domain is the set of positive real numbers and the codomain is a set of real numbers:

$$R(f) = \{x^2/x \in R^+\} = \{1, 4, 9,...\}$$

1.2.1 Types and Applications

Let $f: A \to B$ be a function. f is said to be:

(i) **One-to-One (1–1) function**, if $x_1 \neq x_2$ then $f(x_1) \neq f(x_2)$, $\forall\ x_1, x_2 \in A$.

Or

Whenever $f\{x_1) = f\{x_2)$, $x_1 = x_2$. This function is also known as injective function.

(ii) **Onto (subjective) function**, if for every element y in the codomain B, at least one element x is in the domain A such that $f(x) = Y$.

Or

if $R(f)$ = codomain B.

(iii) **Bijective function**, if f is both 1–1 and onto function.

(iv) **Constant function**, if every element of the domain is mapped to a unique element of the domain (or the domain consists of only one element).

(v) **Into function**, if at least one element of the codomain is not mapped by any element of the domain.

(vi) **Identify function**, if $f(x) = x$, $\forall\ x \in B$, in this case $A \leq B$. [Sometimes it is defined as $f: A \to A$ and $f(x) = x$, $\forall\ x \in A$.]

Consider the following examples:

(i) Let $f: R^+ \to R$ be a function defined as $f(x) = 2(x + 2)$ clearly f is 1 – 1 because if $2(x + 2) = 2(y + 2)$

$$\Rightarrow \quad 2x + 4 = 2y + 4$$
$$\Rightarrow \quad 2x = 2y \Rightarrow x = y$$

Hence, f is 1 – 1.

(ii) Define $f: R \to R^+$ by $f(x) = e^x$, $\forall\ x \in R$. Clearly f is 1 – 1. Because if $f(x_1) = f(x_2)$

$$\Rightarrow \quad e^{x1} = e^{x2}$$
$$\Rightarrow e^{x1-x2} = 1$$
$$\Rightarrow x_1 - x_2 = 0$$
$$\Rightarrow \quad x_1 = x_2$$

Hence, f is 1 – 1.

(iii) Let $A = \{5, 6, 7\}$ and $B = \{a, b\}$, then the mapping $f: A \to B$ is defined as $f(5) = a$; $f(6) = b$: $f(7) = a$. Clearly, f is not 1 – 1. But f is onto.

(iv) Consider Example (ii). If $f: R \to R$, defined by $f(x) = e^x$ is onto. Let x be any element IR^+, then $\log y \in IR$ such that $f\{\log y) = e^{\log y} = y$.

(v) Define $f: Z_+ \to Z_+$ as $f(n) = n^2$, $\forall\ n \in Z_+$. Clearly, f is an into mapping (for example, 3 is not mapped by any element of Z_+) and 1 – 1 mapping but f is not onto.

(vi) Define $f: Z \to Z$ by $f(n) = n + 1$, $\forall\ n \in z$. Clearly, f is 1 – 1 and onto For if (i) $f(n) = f(m) + 1 \Rightarrow n = m\ \therefore\ f$ is 1 – 1.

NOTES

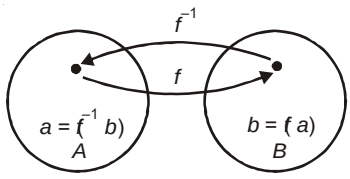
NOTES

(vii) If n is any element of Z, then $n - 1 \in Z$ such that $f(n - 1) = n - 1 + 1 = n$. Hence, f is onto.

Note: A one-one mapping of a set S onto itself is sometimes called a **permutation** of the set S.

Inverse functions and compositions of function

Let f be a bijective function from the set A to the set B. The inverse function of f is the function that assigns to an element $b \in B$ the unique element a in A such that $f(a) = b$. The inverse function of f is denoted as f^{-1}. Hence, $f^{-1}(b) = a$ when $f(a) = b$.



The function f^{-1} is the inverse function of f.

Note: A bijective function is called invertible since we can define an inverse of this function.

Example 1.1:

(i) Define f: $Z \to Z$ by $f(n) = n + 1$. Is f invertible? If it is, what is its inverse?

Solution: The function f has an inverse since it is a bijective function [refer to example (iv)]. Let y be the image of x, so that $y = x + 1$. Then $x = y - 1$, i.e. $y - 1$ is the unique element of Z that is sent to y by f. Hence $f^{-1} = y^{-1}$.

(ii) Let $A = \{a, b, c\}$ and $B = \{5, 6, 7\}$. Define $F: A \to B$ as $f(a) = 5$; $f(b) = 6$; $f(c) = 7$. Is f invertible? If it is, what is its inverse?

Solution: Clearly, the given function is bijective. The inverse function f^{-1} of f is given as $f^{-1}(5) = a$; $f^{-1}(6) = b$; $f^{-1}(7) = c$.

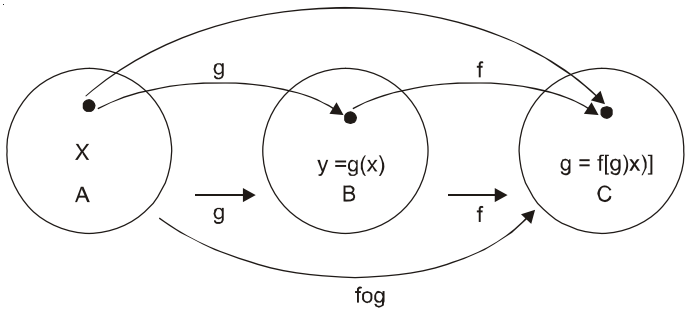
(iii) Define f: $Z \to Z$ by $f(x) = x^2$. ls f invertible?

Solution: Since $f(-2) = f(2) = 4$, f is not 1–1. If an inverse function were defined, it would have to assign two elements to 2. Hence, f is not invertible.

Definition Let g be a function from the set A to the set B and let f be a function from the set B to the set C. The composition of the functions f and g denoted by $(f \circ g)$ is given as follows:

$$(f \circ g)(x) = f(g(x)), \quad \forall\ x \in A$$

NOTES



Example 1.2: Let f: $Z \to Z$ be a function defined by $f(x) = 2x + 3$. Let g: $Z \to Z$ be a function defined by $g(x) = 3x + 2$. Find (i) $f \circ g$ (ii) $g \circ f$.

Solution: Both $f \circ g$ and $g \circ f$ are defined. Further,

(i) $(f \circ g)(x) = f(g(x)) = f(3x + 2)$

$= 2(3x + 2) + 3 = 6x + 7$

(ii) $(g \circ f)(x) = g(f(x)) = g(2x + 3)$

$= 3(2x + 3) + 2 = 6x + 11$

Even though $f \circ g$ and $g \circ f$ are defined, $f \circ g$ and $g \circ f$ need not be equal, i.e. the commutative law does not hold for the composition of functions.

Example 1.3: Let $A = \{1, 2, 3\}$, $B = \{x, y\}$, $C = \{a\}$. Let f:$A \to B$ be defined by $f(1) = x$; $f(2) = y$; $f(3) = x$. Let g: $B \to C$ be defined by $g(x) = a$; $g(y) = a$. Find (i) $f \circ g$, if possible (ii) $g \circ f$, if possible.

Solution:

(i) $(f \circ g)(x) = f(g(x))$ (by definition), but f cannot be applied on C and, hence, $f \circ g$ is meaningless.

(ii) $(g \circ f) : A \to C$ is meaningful. Now $(g \circ f)(x) = g(f(x))$, $\forall\ x \in A$

Hence, $(g \circ f)(1) = g(f(1)) = g(x) = a$

$(g \circ f)(2) = g(f(2)) = g(y) = a$

$(g \circ f)(3) = g(f(3)) = g(x) = a$

Result If f: $A \to B$, g: $B \to C$ and b: $C \to D$, then $(h \circ g)$ of $= h \circ (g \circ f)$.

Proof By the definition $[(h \circ g) \circ f](x) = (h \circ g)(f(x))$

$= h[(g(f(x)]$ (i)

and $[h \circ (g \circ f)](x) = h[(g \circ f)(x)]$

$= h[g(f(x))]$ (ii)

From (i) and (ii), $(h \circ g)$ of $= h \circ (g \circ f)$

Result Let f: $A \to B$ and g : $B \to C$. Then

(i) $g \circ f$ is onto if both f and g are onto.

(ii) $g \circ f$ is $1 - 1$ if both f and g are $1 - 1$.

NOTES

Proof

(i) Let $Z \in C$. Since $g : B \to C$ is onto, F an element $y \in B$ such $g(y) = Z$. Since $f: A \to B$ is onto, for an element $x \in B$, such that $f(x) = y$

$\therefore (g \ o \ f)(x) = g(f(x)) = g(y) = Z$

$\therefore (g \ o \ f)$ is onto.

(ii) Let $x_1 \neq x_2$ be two elements in A. Since $f: A \to B$ is $1 - 1$ and $f(x_1) \neq (x_2)$, $g(f(x_1)) = g(f(x_2))$. Thus, $g \ o \ f$ is $1 - 1$. In B, $g: B \to C$ is $1 - 1$ and $f(x_1) \neq (x_2)$.

Some important functions

Greatest integer function: The greatest floor function assigns to the real number x the largest integer that is less than or equal to x, and the value of this function is denoted by $[x]$ (or $\lfloor x \rfloor$)

Ceiling: The ceiling function assigns to the real number x the smallest integer that is greater than or equal to x. The value of this function is denoted by $\lceil x \rceil$.

Consider the following examples:

(i) $\lfloor x \rfloor = \lfloor 1/2 \rfloor = 0; \lceil x \rceil = \lceil 1/2 \rceil = 1$

(ii) $\lfloor 5.6 \rfloor = 5; \lfloor 5.6 \rfloor = 6$

(iii) $\lfloor 4.1 \rfloor = 4; \lfloor 4.1 \rfloor = 5$

(iv) $\lfloor 3 \rfloor = 3; \lfloor 3 \rfloor = 3$

Example 1.4: Data stored on a computer disk or transmitted over a data network is represented as a string of bytes. How many bytes are required to encode 500 bits of data?

Solution: To find the number of bytes needed, we determine the smallest integer that is at least as large as the quotient. When 500 is divided by 8, the number of bits in a byte is:

$$\left\lceil \frac{500}{8} \right\rceil = \lceil 62.5 \rceil = 63 \text{ bytes is required}$$

Example 1.5: In ATM (asynchronous tranfer mode), data are organized into cells of 53 bytes. How many ATM cells can be transmitted in 2 minutes over a connection that transmits data at the rate of 200 kB per second.

Solution: In 2 minutes, this connection can transmit $200{,}000 \times 60 \times 2 = 2{,}40{,}00{,}000$ bits. Since each ATM cell is 53 bytes ($53 \times 8 = 424$ bits long), the number of ATM cells that can be transmitted in 2 minutes in the given connection is

$$\left[\frac{2{,}40{,}00{,}000}{424} \right]$$

Modulus operator If x is a non-negative integer and y is a positive integer, we define x mod y to be the remainder when x is divided by y. For example,

$$11 \bmod 2 = 1;\ 5 \bmod 1 = 0;\ 365 \bmod 7 = 1$$

NOTES

Example 1.6:

(*i*) What day of the week will it be 365 days from Friday?

(*ii*) What is meant by ISBN 0-07-0035 75-*x*?

Solution:

(*i*) 365 mod 7 = 1. Thus 365 days from Friday, it will be saturday. Since 7 days after Friday is Friday, in general, if $K>0$, $K \in Z$, after $7K$ days it is Friday again.

(*ii*) ISBN is 0-07-003575-*x*. Here 0-stands for the book as one from an English-speaking country. The publisher code 07 identifies the book as one published by McGraw-Hill. The code 003575 uniquely identifies the book as one published by McGraw-Hill. The check character is n mod 11, where

$n = 0 + 2.0 + 3.7 + 4.0 + 5.0 + 6.3 + 7.5 + 8.7 + 9.5$

$= 0 + 0 + 21 + 0 + 0 + 18 + 35 + 56 + 45$

$n = 175$

$\therefore$ n mod 11 = 175 mod 11 = 10

$\therefore$ The check character is 10, i.e. it is represented as *x*.

Hashing function: The central computer (server) at an university maintains records for each student. How many memory locations can be alloted so that the student records can be retrieved fast? The solution to this problem is to use a suitably chosen hasing function. Records are identified using a key, which uniquely identifies each student's record. For example, student records are identified using the register number/enrolment number of the student as the key. A hashing function h assigns memory location $h(p)$ to the record that has p as its key.

The most common hashing function is the function $h(p) = p$ mod q, where q is the number of available memory locations. The hashing function should be onto so that all memory allocations are possible. For example, if $q = 11$, the record of the student with register number 15 is assigned to memory location 4, since $h(15)=4$ mod 11. Similarly, since $h(122) = 122$ mod 11 = 2, the record of the student with register number 121 is assigned to memory location 2.

Since a hashing function is not 1 – 1, more than one file may be assigned to a memory location. When this happens, we say that a collision occurs. To resolve a collision is to assign the first free location following the occupied memory location assigned by the hashing function. For instance, after making two earlier assignments, we assign location 5 to the record of the student with register number 256.

NOTES

CHECK YOUR PROGRESS

1. What is a function in terms of mathematics?
2. Name the various types of functions.
3. What is a ceiling function?
4. Which is the most important hashing function?

1.3 ARITHMETIC PROGRESSION

The arithmetic progression, also referred to as an arithmetic sequence, is a sequence of numbers in a manner in which the difference of any two consecutive members of the sequence is a constant. In other words, Quantities $a_1, a_2, a_3, ..., a_n$ are said to be in *Arithmetical Progression* if $a_n - a_{n-1}$ is constant for all integers $n > 1$. The constant quantity $a_n - a_{n-1}$ is called the *common difference* of the arithmetical progression.

Notation. AP stands for an arithmetical progression. Consider the following series

$$1, 3, 5, 7, 9, 11, ...$$

$$0, \sqrt{2}, 2\sqrt{2}, 3\sqrt{2}, 4\sqrt{2}, ...$$

$$1, \frac{1}{2}, 0, -\frac{1}{2}, -1, -\frac{3}{2}, ...$$

$$x + y, x, x - y, x - 2y, ...$$

$$5.3, 5.55, 5.8, 6.05, 6.3, ...$$

Each of the series is an AP. Common differences are 2, $\sqrt{2}$, $-\frac{1}{2}$, $-y$ and 0.25 respectively.

1.3.1 Finding the General Term of a Given Arithmetical Progression

Let $a_1, a_2, ..., a_n, ...$ be a given AP. Let d be their common difference. Then $a_n - a_{n-1} = d$ for all n.

$\Rightarrow$ $a_2 - a_1 = d$, $a_3 - a_2 = d$, $a_4 - a_3 = d$, and so on,

$\Rightarrow$ $a_2 = a_1 + d$, $a_3 = a_2 + d = a_1 + d + d = a_1 + 2d$

$$a_4 = a_3 + d = a_1 + 2d + d = a_1 + 3d$$

$$... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ...$$

$$a_{n-1} = a_1 + (n-2)d$$

$$a_n = a_{n-1} + d = a_1 + (n-2)d + d$$

$$= a_1 + (n-1)d.$$

Thus, nth term a_n of arithmetical progression whose first term is a_1 and common difference d is given by

$$a_n = a_1 + (n-1)d$$

NOTES

Example 1.7: Find 16th term of the series 3.75, 3.5, 3.25, ...

Solution: In this case $a_1 = 3.75$, $a_2 = 3.5$, $a_3 = 3.25$

$$d = a_2 - a_1 = -.25$$

Hence, 16th term = a_{16}

$$= 3.75 + (16 - 1)(-.25)$$
$$= 3.75 - 15 \times .25$$
$$= 3.75 - 3.75 = 0.$$

Example 1.8: Which term of the AP 49, 44, 39, ... is 9?

Solution: Let nth term be 9, i.e. $a_n = 9$.

Here $a_1 = 49$, $d = 44 - 49 = -5$,

Thus $a_n = 9 = 49 + (n - 1)(-5)$

$\Rightarrow$ $9 = 49 - 5n + 5$

or $5n = 54 - 9 = 45$

$n = 9$

Thus 9th term of the given AP is 9.

1.3.2 Finding the Sum of a Finite Number of Quantities

Let $a_1, a_2, ..., a_n$ be n quantities in AP, and let the last term a_n be denoted by l. If d is their common difference then

$$a_n = a_1 + (n - 1)d = l.$$

Put $S_n = a_1 + a_2 + \cdots + a_n$

Thus $S_n = a_1 + (a_1 + d) + (a_2 + 2d) + \cdots + [a_1 + (n - 1)d]$

$$= a_1 + (a_1 + d) + (a_1 + 2d) + \cdots + (l - d) + l \qquad (1.1)$$

Writing, the above series in reverse order, we get:

$$S_n = l + (l - d) + (l - 2d) + \cdots + (a_1 + d) + a_1 \qquad (1.2)$$

Adding equations (1.1) and (1.2), we get:

$$2S_n = (a_1 + l) + (a_1 + l) + \cdots + (a_1 + l), (n \text{ times})$$
$$= n(a_1 + l)$$

Therefore, $S_n = \frac{n}{2}(a_1 + l)$

$$= \frac{n}{2}\{a_1 + [a_1 + (n - 1)d]\}$$

Consequently $S_n = \frac{n}{2}[2a_1 + (n - 1)d]$.

Example 1.9: Find the sum of $\frac{3}{4}, \frac{2}{3}, \frac{7}{12}, ...$ up to 19 terms.

Solution: Here $a_1 = \frac{3}{4}$, $a_2 = \frac{2}{3}$, $n = 19$.

Thus $d = a_2 - a_1 = \frac{2}{3} - \frac{3}{4} = -\frac{1}{12}$

NOTES

i.e. $$S_{10} = \frac{19}{2}\left[2\times\frac{3}{4}+(19-1)\left(-\frac{1}{12}\right)\right]$$

$$= \frac{19}{2}\left[\frac{3}{2}-\frac{18}{12}\right]$$

$$= \frac{19}{2}\left(\frac{18-18}{12}\right) = \frac{19}{2}\times 0 = 0$$

Example 1.10: How many terms of the following series may be taken so that their sum is 66:

$$-9, -6, 3, \ldots ?$$

Solution: Let $S_n = 66$. Here $a_1 = -9$, $d = -6 + 9 = 3$.

or $$66 = \frac{n}{2}[-18+(n-1)3]$$

$$132 = -18n + 3n^2 - 3n$$

i.e. $$3n^2 - 21n - 132 = 0$$

or $$n^2 - 7n - 44 = 0 \Rightarrow (n-11)(n+4) = 0$$

$$\Rightarrow n = 11 \text{ or } n = -4$$

As n is a positive integer, the second value, –4, is rejected.

Thus, the required number of terms = 11.

1.3.3 Arithmetic Mean

Simply speaking, the arithmetic mean of a list of numbers refers to the total of all of the list divided by the number of items in the list. If $a_1, a_2, \ldots, a_n$ are in AP then the quantities $a_2, a_3, \ldots, a_{n-1}$ are called arithmetic means (written as AM) between a_1 and a_n.

Thus, in the series, 1, 3, 5, 7, 9, 11, 13, 15, ...

3, 5 are arithmetic means between 1 and 7 and 9, 11, 13 are arithmetic means between 7 and 15.

To insert *n* arithmetic means between two given quantities

Let a and b be two given quantities and $A_1, A_2, \ldots, A_n$ be the n arithmetic means between them. Then the quantities

$$a, A_1, A_2, \ldots, A_n, b \text{ are in A. P.}$$

Let d be their common difference.

Now $$b = (n+2)\text{th term}$$

$$= a + (n+1)d$$

$$\Rightarrow \quad d = \frac{b-a}{n+1}$$

Further, $$A_1 = 2\text{nd term}$$

$$= a + d = a + \left(\frac{b-a}{n+1}\right)$$

$$= \frac{na+b}{n+1}$$

$$A_2 = 3\text{rd term}$$

$$= a + 2d = a + 2\left(\frac{b-a}{n+1}\right)$$

$$= \frac{na+2b-a}{n+1}$$

... ...

$$A_n = a + nd = a + n\left(\frac{b-a}{n+1}\right)$$

$$= \frac{a+nb}{n+1}.$$

Hence, $\frac{na+b}{n+1}, \frac{na+2b-a}{n+1}, ..., \frac{a+nb}{n+1}$

are n arithmetic means between a and b.

Example 1.11: Insert 6 arithmetic means between 1 and 19.

Solution: Let $A_1, A_2, A_3, A_4, A_5, A_6$, be the required arithmetic means.

Then 1, $A_1, A_2, A_3, A_4, A_5, A_6$, 19 are in AP.

Let d be their common difference.

Then, $19 = 8\text{th term}$

$$= 1 + (8-1)d$$

$$= 1 + 7d$$

Thus, $d = \frac{18}{7}$

Hence, $A_1 = 2\text{nd term}$

$$= 1 + \frac{18}{7} = \frac{25}{7}$$

$$A_2 = 3\text{rd term} = 1 + 2 \times \frac{18}{7} = 1 + \frac{36}{7} = \frac{43}{7}$$

$$A_3 = 4\text{th term} = 1 + 3 \times \frac{18}{7} = 1 + \frac{54}{7} = \frac{61}{7}$$

$$A_4 = 5\text{th term} = 1 + 4 \times \frac{18}{7} = 1 + \frac{72}{7} = \frac{79}{7}$$

$$A_5 = 6\text{th term} = 1 + 5 \times \frac{18}{7} = 1 + \frac{90}{7} = \frac{97}{7}$$

$$A_6 = 7\text{th term} = 1 + 6 \times \frac{18}{7} = 1 + \frac{108}{7} = \frac{115}{7}$$

So, the required means are:

$$\frac{25}{7}, \frac{43}{7}, \frac{61}{7}, \frac{79}{7}, \frac{97}{7}, \frac{115}{7}$$

NOTES

Example 1.12: How many terms of the series 4, 2, 0, –2, –4, –6,... may be taken so that their sum is 6?

Solution: Let n be the required number of terms.

Here $a_1 = 4$, $d = 2 - 4 = -2$

So $$6 = S_n = \frac{n}{2}[8 + (n-1)(-2)]$$

i.e. $$12 = 8n - 2n^2 + 2n$$

or $$2n^2 - 10n + 12 = 0$$

or $$n^2 - 5n + 6 = 0$$

or $$(n-3)(n-2) = 0$$

Hence, $$n = 3 \text{ or } 2$$

We get two answers. As the third term is zero, therefore, the sum of the first two terms is the same as that of the first three terms.

Example 1.13: If pth, qth, rth term of an AP are a, b, c, respectively, show that $(q-r)\,a + (r-p)b + (p-q)c = 0$.

Solution: Here pth term $= a = a_1 + (p-1)d$ (i)

qth term $= b = a_1 + (q-1)d$ (ii)

rth term $= c = a_1 + (r-1)d$ (iii)

where a_1 is the first term and d is the common difference of the AP.

Multiply (i) by $q - r$, (ii) by $r - p$, (iii) by $p - q$ and add to obtain $(q-r)\,a + (r-p)\,b + (p-q)\,c$

$$= a_1(q-r) + a_1(r-p) + a_1(p-q) + d[(p-1)(q-r) + (q-1)(r-p) + (r-1)(p-q)]$$

$$= a_1[q - r + r - p + p - q] + d[pq + r - pr - q + qr - r + p - pq + rp - rq - p + q]$$

$$= 0$$

Example 1.14: The sum of n terms of two APs are in the ratio of $7n + 1$: $4n + 27$. Find the ratio of their 11th terms.

Solution: Let a_1 and b_1 be the first terms of two APs and d_1, d_2 be their common difference respectively.

Then, $$S_n = \frac{n}{2}[2a_1 + (n-1)\,d_1]$$

$$S'_n = \frac{n}{2}[2b_1 + (n-1)\,d_2]$$

So, $$\frac{S_n}{S'_n} = \frac{2a_1 + (n-1)d_1}{2b_1 + (n-1)d_2} = \frac{7n+1}{4n+27}$$

Putting $n = 21$, we get

$$\frac{2a_1 + 20d_1}{2b_1 + 20d_2} = \frac{148}{111}$$

or $$\frac{a_1 + 10d_1}{b_1 + 10d_2} = \frac{148}{111}$$

NOTES

or $$\frac{a_{11}}{b_{11}} = \frac{148}{111}$$

where a_{11} and b_{11} are 11th terms of two APs respectively.

$$= \frac{4}{3}$$

Example 1.15: If a^2, b^2, c^2 are in AP, show that

$$\frac{1}{b+c}, \frac{1}{c+a}, \frac{1}{a+b} \text{ are also in AP.}$$

Solution: $\frac{1}{b+c}, \frac{1}{c+a}, \frac{1}{a+b}$ are in AP.

$$\Leftrightarrow \quad \frac{1}{c+a} - \frac{1}{b+c} = \frac{1}{a+b} - \frac{1}{c+a}$$

$$\Leftrightarrow \quad \frac{b-a}{(c+a)(b+c)} = \frac{c-b}{(a+b)(c+a)}$$

$$\Leftrightarrow \quad \frac{b-a}{c+b} = \frac{c-b}{b+a}$$

$$\Leftrightarrow \quad b^2 - a^2 = c^2 - b^2$$

$\Leftrightarrow$ a^2, b^2, c^2 are in AP.

Hence, the result follows.

Example 1.16: Sum to n terms of three APs are s_1, s_2 and s_3. The first term of each of them is 1 and common differences are 1, 2 and 3 respectively. Show that s_1, s_2, s_3 are also in AP.

Solution: $$s_1 = \frac{n}{2}[2 + (n-1)] = \frac{n(n+1)}{2}$$

$$s_2 = \frac{n}{2}[2 + (n-1)2] = n^2$$

$$s_3 = \frac{n}{2}[2 + (n-1)3] = \frac{n}{2}(3n-1)$$

or $$s_2 - s_1 = n^2 - \frac{n(n+1)}{2} = \frac{n^2 - n}{2} = \frac{n(n-1)}{2}$$

$$s_3 - s_2 = \frac{n}{2}(3n-1) - n^2 = \frac{3n^2 - n - 2n^2}{2}$$

$$= \frac{n^2 - n}{2} = \frac{n(n-1)}{2}$$

i.e. $s_2 - s_1 = s_3 - s_2$

Hence, s_1, s_2, s_3 are also in AP.

Example 1.17: State giving example whether the following statement is true or false.

In a given AP let a be the first term, d the common difference, n the number of terms and s their sum. Given any three of a, d, n and s one can always find a unique value of the fourth quantity.

NOTES

Solution: The statement is false. Given the values of a, d and s, there might be two values of n. Consider the following series:

$$12, 9, 6, 3, 0, -3, -6$$

It is to find the number of terms from beginning whose sum is 30.

In this case, $a = 12$, $d = 9 - 12 = -3$, $s = 3$

Since $$s = \frac{n}{2}[2a + (n-1)d]$$

we have $$30 = \frac{n}{2}[24 + (n-1)(-3)]$$

or $$60 = 24n - 3n^2 + 3n$$

or $$3n^2 - 27n + 60 = 0$$

$$\Rightarrow \quad n^2 - 9n + 20 = 0$$

$$\Rightarrow \quad (n-4)(n-5) = 0$$

$$\Rightarrow \quad n = 4 \text{ or } 5$$

Example 1.18: Find the sum of all the numbers between 200 and 400 which are divisible by 7.

Solution: Since 4 is left as remainder on division of 200 by 7, the least number greater than 200 divisible by 7 is 203. Again, if we divide 400 by 7, 1 is left as remainder. This implies that the greatest number less than 400, which is divisible by 7 is 399.

So, we are to find the sum of the series

$203 + 210 + 217 + \cdots + 399$

Here, $a_1 = 203$, $d = 7$, $l = 399$

Let n be the total number of terms in this series.

Then, $$399 = 203 + (n-1)7$$

$$\Rightarrow \quad 7n = 399 - 203 + 7$$

$$= 406 - 203 = 203$$

$$\Rightarrow \quad n = 29.$$

Hence, required sum $= \frac{n}{2}(a + l) = \frac{29}{2}(203 + 399)$

$$= \frac{29}{2}(602)$$

$$= (29)(301)$$

$$= 8729$$

Example 1.19: Mr X arranges to pay off a debt of Rs 9,600 in 48 annual instalments which form an arithmetical series. When 40 of these instalments are paid, Mr X becomes insolvent and his creditor finds that Rs 2,400 still remains unpaid. Find the value of each of the first three instalments of Mr X. Ignore the interest.

Solution: Let a, $a + d$, $a + 2d$, $a + 3d$ be the annual instalments.

The sum of this series up to $n = 48$ terms is 9600.

i.e. $$9600 = \frac{48}{2}\,[2a + (48 - 1)d]$$

$\Rightarrow$ $$9600 = 24(2a + 47d)$$

$\Rightarrow$ $$2a + 47d = 400 \qquad \text{(i)}$$

After 40 instalments are paid, the balance is Rs 2400. In other words, in 40 instalments Mr X has paid Rs (9600 – 2400) i.e. Rs 7200.

So, the sum of first 40 terms of the above series is 7200.

Thus $$7200 = \frac{40}{2}\,[2a + (40 - 1)d]$$

$\Rightarrow$ $$7200 = 20(2a + 39d)$$

$\Rightarrow$ $$2a + 39d = 360 \qquad \text{(ii)}$$

Subtracting (ii) from (i), we get:

$$8d = 40 \quad \Rightarrow \quad d = 5$$

Then (2) $\Rightarrow 2a + 195 = 360$

$\Rightarrow$ $$2a = 165 \quad \Rightarrow \quad a = 82.50$$

Hence, first instalment of Mr X is Rs 82.50, second instalment is Rs (82.50 + 5.00), i.e. Rs 87.50 and third instalment is Rs (87.50 + 5.00), i.e. Rs 92.50.

Example 1.20: If $\frac{1}{b+c}, \frac{1}{c+a}, \frac{1}{a+b}$ are in AP, prove that a^2, b^2 c^2 are also in AP.

Solution: Since $\frac{1}{b+c}, \frac{1}{c+a}, \frac{1}{a+b}$ are in AP.

We have $$\frac{1}{c+a} - \frac{1}{b+c} = \frac{1}{a+b} - \frac{1}{c+a}$$

$\Rightarrow$ $$\frac{(b+c)-(c+a)}{(b+c)(c+a)} = \frac{(c+a)-(a+b)}{(a+b)(c+a)}$$

$\Rightarrow$ $$\frac{b-a}{b+c} = \frac{c-b}{b+a}$$

$\Rightarrow$ $$b^2 - a^2 = c^2 - b^2$$

$\Rightarrow a^2, b^2, c^2$ are in AP.

Note: This question was also solved implicitly in Example 1.15.

Example 1.21: The monthly salary of a person was Rs 320 for each of the first three years. He next got annual increments of Rs 40 per month for each of the following successive 12 years. His salary remained stationary till retirement when he found that his average monthly salary during the service period was Rs 698. Find the period of his service.

NOTES

Solution: Let n be the total number of year of the person's service.

His total salary = Rs $12n \times 698$

(As his monthly average salary is Rs 698)

Total salary in first three years of service

$= 320 \times 3 \times 12 =$ Rs 960×12

In the 4th year, his monthly salary was Rs (320 + 40) = Rs 360; in the 5th year his monthly salary was Rs 400, and so on.

Then, for the next 12 years, his total salary

$= \text{Rs } 12 \times [360 + 400 + \cdots \text{ up to 12 terms}]$

$= \text{Rs } 12 \times \frac{12}{2}[2 \times 360 + (12 - 1) .40]$

$= \text{Rs } 12 \times 6(720 + 440)$

$= \text{Rs } 12 \times 6 \times 1160$

$= \text{Rs } 12 \times 6960$

At the end of following 12 years, his monthly salary was

Rs $[360 + (12 - 1)40] =$ Rs 800

This salary he got for the remaining $(n - 15)$ years. So, his total salary for the remaining $(n - 15)$ years was $(n - 15)\ 800 \times 12$.

Hence, his total salary throughout his service period

$= 12[960 + 6960 + 800(n - 15)]$

$= 12(7920 + 800n - 12000)$

$= 12\ (800n - 4080)$

This must be the same as $12n \times 698$

i.e. $12n \times 698 = 12(800n - 4080)$

$\Rightarrow \quad 102n = 4080 \quad \Rightarrow \quad n = 40$ years

Example 1.22: The sequence of natural numbers is written as:

		1		
	2	3	4	
5	6	7	8	9
...	...	...	...	...
...	...	...	...	...

Find the sum of the numbers in the rth row.

Solution: Let us consider that S_1 denotes the sum of rth row.

$S_1 = 1,\ S_2 = 2 + 3 + 4,\ S_3 = 5 + 6 + 7 + 8 + 9$

Let initial term of S_k be t_k

and suppose that $M = 1 + 2 + 5 + \cdots + t_k$ be the sum of first terms of S_1, S_2, and S_k.

Now, $M = 1 + 2 + 5 + 10 + \cdots + t_k$

Also, $M = 1 + 2 + 5 + \cdots + t_{k-1} + t_k$

Subtracting, we get:

$$0 = (1 + 1 + 3 + 5 + \text{up to } k \text{ term}) - t_k$$

$$t_k = 1 + [1 + 3 + 5 + \cdots \text{ up to } (k-1) \text{ terms}]$$

$$= 1 + \left(\frac{k-1}{2}\right)[2 + (k-2).2]$$

$$= 1 + (k-1)^2 = k^2 - 2k + 2$$

Also, in S_1 there is one term, in S_2 there are 3 terms, and so on. Hence, in S_k there will be $(2k-1)$ term.

Hence, we are to find the sum of series

$r^2 - 2r + 2$, $r^2 - 2r + 3$, $r^2 - 2r + 4$, up to $(2r-1)$ terms

So
$$S_k = \frac{(2r-1)}{2}\ [2(r^2 - 2r + 2) - (2r-2).1]$$

$$= (2r-1)\ (r^2 - 2r + 2 + r - 1)$$

$$= (2r-1)\ (r^2 - r + 1)$$

$$= 2r^3 - 3r^2 + 3r - 1$$

Example 1.23: A lamp lighter has to light 100 gas lamps. He takes 1½ minutes to go from one lamp post to the next. Each lamp post burns 10 cc of gas per hour. How many cc of gas has been burnt by 8.30 PM if he lights the first lamp at 6 PM.

Solution: Total time from 6 PM to 8.30 PM is 2.30 hours, i.e. 150 minutes.

g_1 = Gas burnt in first lamp = $\frac{150 \times 10}{60}$

$$= \frac{1}{6}(150) \text{ cc}$$

Second lamp burns for $(150 - 1.5)$ minutes.

So, the gas burnt in second lamp = $g_2 = \frac{1}{6}\ (150 - 1.5)$

Similarly, $g_3 = \frac{1}{6}\ (150 - 2 \times 1.5)$, and so on.

We are to calculate

$$S = g_1 + g_2 + \cdots + g_{100}$$

$$S = \frac{1}{6}[150 + (150 - 1.5) + (150 - 2 \times 1.5) + \cdots]$$

The series inside bracket is an AP with first term = 150, common difference = –1.5 and number of terms = 100

Hence
$$S = \frac{1}{6}\frac{100}{2}[300 + 99(-1.5)]$$

$$= \frac{25}{3}\ (300 - 1.5 \times 99)$$

$$= 25\ (100 - 1.5 \times 33)$$

$= 25 (100 – 49.5)$

$= 25 (50.5)$

$= 1265.255$ cc.

NOTES

Example 1.24: Two posts were offered to a man. In one the starting salary was Rs 120 per month and the annual increment was Rs 8; in the other post the salary commenced at Rs 85 per month but the annual increment was Rs 12. The man decided to accept the post which would give him more earnings in the first twenty years of the service. Which post was acceptable to him? Justify your answer.

Solution: The total earnings of the man in first job:

$$= \frac{20}{2}[2 \times 120 + (20 – 1)8] \times 12$$

$= 10 (240 + 152) \times 12$

$= 120 (392)$

$=$ Rs 47,040

His total earnings in second job:

$$= \frac{20}{2}[2 \times 85 + (20 – 1)12] \times 12$$

$= 120 (170 + 228)$

$= 120 (398)$

$=$ Rs 47,760

which is greater than Rs 47,040. Hence, the second job was acceptable to the man.

Example 1.25: Find the sum of all natural numbers between 500 and 1000 which are divisible by 13.

Solution: 500 leaves, on division by 13, 6 as remainder, so 507 is the least number greater than 500 which is divisible by 13.

Also, the remainder, when 1000 is divided by 13 is 12. So, the greatest number less than 1000, divisible by 13 is 988.

We are to find the sum $507 + 520 + \cdots + 988$.

Let n be the number of terms in the series.

Then, $988 = 507 + (n – 1)\ 13$

$\Rightarrow$ $76 = 39 + n – 1 = n + 38$

$\Rightarrow$ $n = 76 – 38 = 38$

Required sum, $S = \frac{n}{2}[2a + (n – 1)d]$

Here, $n = 38,\ a = 507,\ d = 13$

Hence, $S = \frac{38}{2}\ [2 \times 507 + (38 – 1)13]$

$$= \frac{38}{2}\ (1014 + 37 \times 13)$$

$= 19\ (1014 + 481)$

$= 19\ (1495)$

$= 28{,}405$

Example 1.26: A firm produced 1000 sets of TV during its first year. The total sum of the firm's production at the end of 10 year's operation is 14,500 sets.

(i) Estimate by how many units, the production increased each year, if the increase in each year is uniform.

(ii) Forecast, based on the estimate of the annual increment in production, the level of output for the 15th year.

Solution: Here $a = 1000$, $S = 14{,}500$, $n = 10$ and d is to be evaluated.

$$14{,}500 = \frac{10}{2}[2 \times 1000 + (10 - 1)d]$$

$$14500 = 5\ (2000 + 9d)$$

$$\Rightarrow \quad 2000 + 9d = 2900$$

$$\Rightarrow \quad 9d = 900 \Rightarrow \quad d = 100$$

Hence, 1000 units is the per annum increase

$$a + 14d = 1000 + 14 \times 100$$

$$= 1000 + 1400 = 2400$$

CHECK YOUR PROGRESS

5. Define arithmetic progression.
6. What do you understand by the arithmetic mean of a list of numbers?

1.4 GEOMETRIC PROGRESSION

A geometric progression refers to a sequence of numbers in which each number is obtained from the previous one by multiplying it by a constant. Non-zero quantities $a_1, a_2, a_3, ..., a_n$ each term of which is equal to the product of the preceding term and a constant number are called to form a Geometrical Progression (written as GP).

Thus, all the following quantities are in GP.

(*i*) 1, 2, 4, 8, 16,...

(*ii*) $3, -1, \frac{1}{3}, \frac{-1}{9}, \frac{1}{27}, ...$

(*iii*) $1, \sqrt{2}, 2, 2\sqrt{2}, ...$

(*iv*) $a, \frac{a}{b}, \frac{a}{b^2}, \frac{a}{b^3}, ...$, where $a \neq 0, b \neq 0$.

(*v*) $1, \frac{1}{5}, \frac{1}{25}, \frac{1}{125}, ...$

The constant number is termed as the *common ratio* of the GP.

NOTES

1.4.1 Finding the *n*th Term of a Geometric Progression

Let first term be a and r the common ratio, By definition the GP is $a, ar, ar^2, \ldots$

$$1\text{st term} = a = ar^0 = ar^{1-1}$$

$$2\text{nd term} = ar = ar^1 = ar^{2-1}$$

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

In general, nth term $= ar^{n-1}$.

In preceding examples, we compute 5th, 7th, 3rd, 11th and 8th term of (*i*), (*ii*), (*iii*), (*iv*) and (*v*), respectively.

In (*i*), Ist term is 1 and common ratio = 2

Hence, 5th term $= ar^4 = 1.2^4 = 16$

In (*ii*), $a = 3$ and $r = \frac{-1}{3}$

Hence, 7th term $= ar^6 = 3\left(\frac{-1}{3}\right)^6 = \frac{1}{243}$

In (*iii*), $a = 1$, $r = \sqrt{2}$

Hence, 3rd term $= ar^2 = 2$

In (*iv*), Ist term $= a$, $r = \frac{1}{b}$

Hence, 11th term $= ar^{10} = \frac{a}{b^{10}}$

In (*v*), $a = 1$, $r = \frac{1}{5}$

Hence, 8th term $= ar^2 = \frac{1}{5^7} = \frac{1}{78{,}125}$

1.4.2 Finding the Sum of First *n* Terms of a Geometric Progression

Let $a, ar, ar^2, \ldots$ be a given GP and let S_n be the sum of its first n terms.

Then, $S_n = a + ar + ar^2 + \cdots + ar^{n-1}$.

This gives $rS_n = ar + ar^2 + \cdots + ar^{n-1} + ar^n$

Subtracting, we get:

$$S_n - r\,S_n = a - ar^n = a\,(1 - r^n)$$

In case $r \neq 1$, $S_n = \frac{a\left(1-r^n\right)}{(1-r)}$

In case $r = 1$, $S_n = a + a + a + \cdots + a$ (n times)

$= na$.

Thus, sum of n terms of a GP is $\frac{a\left(1-r^n\right)}{1-r}$ provided $r \neq 1$.

In case $r = 1$, sum of GP is na.

NOTES

Example 1.27: Find the sum of the first 14 terms of a GP

3, 9, 27, 81, 243, 729,...

Solution: In this case $a = 3$, $r = 3$, $n = 14$

$$\text{So,} \quad S_n = \frac{a(1-r^n)}{1-r} = \frac{3(1-3^{14})}{1-3}$$

$$= \frac{3}{2}(3^{14} - 1)$$

Example 1.28: Find the sum of first 11 terms of a GP given by

$$1, -\frac{1}{2}, \frac{1}{4}, -\frac{1}{8} \ldots$$

Solution: Here $a = 1$, $r = -\frac{1}{2}$, $n = 11$

$$\text{So,} \quad S_n = \frac{a(1-r^n)}{1-r} = \frac{1\left[1-\left(-\frac{1}{2}\right)^{11}\right]}{1+\frac{1}{2}}$$

$$= \frac{2^{11}+1}{3\times 2^{10}} = \frac{683}{1024}$$

1.4.3 Finding the Sum to Infinity of a Geometric Progression Whose Common Ratio is Less Than 1

Let $a, ar, ar^3, \ldots$ be a GP with $r<1$.

Now $r < 1 \quad \Rightarrow \quad r^2 < r, r^3 < r^2, \ldots$.

Thus, as power of r goes on increasing, the corresponding term in GP decreases in value. So, we can assume that as n becomes indefinitely large, r^n becomes indefinitely small, i.e. $r^n \to 0$.

$$\text{Now} \qquad S_n = \frac{a(1-r^n)}{1-r} = \frac{a}{1-r} - \frac{ar^n}{1-r}$$

So, as $n \to \infty$, $S_\infty = \frac{a}{1-r}$

Example 1.29: Find the sum of the following series up to infinity

$$1 + \frac{3}{7} + \frac{9}{49} + \frac{27}{343} + \frac{81}{2401} + \cdots$$

Solution: Here $a = 1$, $r = \frac{3}{7} < 1$.

$$\text{So,} \quad S_\infty = \frac{a}{1-r} = \frac{1}{1-\frac{3}{7}} = \frac{7}{4}$$

NOTES

Example 1.30: Evaluate the recurring decimal 17.

Solution: $0.17 = 0.1 + 0.07 + 0.007 + 0.0007 + \cdots$

$$= \frac{1}{10} + \frac{7}{10^2} + \frac{7}{10^3} + \cdots$$

$$= \frac{1}{10} + \frac{7}{10^2}\left\{1 + \frac{1}{10} + \frac{1}{10^2} + \cdots\right\}$$

$$= \frac{1}{10} + \frac{7}{10^2}\frac{1}{1 - \frac{1}{10}}$$

$$= \frac{1}{10} + \frac{7}{100}\frac{10}{9}$$

$$= \frac{1}{10} + \frac{7}{90} = \frac{16}{90} = \frac{8}{45}$$

1.4.4 Geometric Mean

The geometric mean refers to a type of mean or average which indicates the central tendency or typical value of a set of numbers. If α, β and γ are in GP then β is called a Geometric Mean (GM) between α and γ.

If a_1, a_2, ..., a_n are in GP, then a_2, ..., a_{n-1} are called Geometric Means between a_1 and a_n.

Thus, 3, 9, 27 are three geometric means between 1 and 81.

To find *n* GMs between two given numbers, *a* and *b*

Let G_1, G_3, ..., G_n be n GMs between a and b. Thus, a, G_1, G_2, ..., G_n, b is a GP, b being $(n + 2)$th term $= ar^{n+1}$, where r is the common ratio of GP.

Thus, $b = ar^{n+1} \Rightarrow r = \left(\frac{b}{a}\right)^{\frac{1}{n+1}}$

So, $G_1 = ar = a\left(\frac{b}{a}\right)^{\frac{1}{n+1}} = \left(a^n b\right)^{\frac{1}{n+1}}$

$G_2 = ar^2 = a\left(\frac{b}{a}\right)^{\frac{2}{n+1}} = \left(a^{n-1}b^2\right)^{\frac{1}{n+1}}$

... .

$G_n = ar^{n-1} = a\left(\frac{b}{a}\right)^{\frac{n-1}{n+1}} = \left(a^2 b^{n-1}\right)^{\frac{1}{n+1}}$

Example 1.31: Find 7 GMs between 1 and 256.

Solution: Let G_1, G_2, ..., G_7, be 7 GMs between 1 and 256.

Then, 256= 9th term of GP,

$= 1\ r^8$ where r is the common ratio of the GP

NOTES

This gives:

$$r^8 = 256 \Rightarrow r = 2$$

Thus
$$G_1 = ar = 1 \times 2 = 2$$
$$G_2 = ar^2 = 1 \times 4 = 4$$
$$G_3 = ar^3 = 1 \times 8 = 8$$
$$G_4 = ar^4 = 1 \times 16 = 16$$
$$G_5 = ar^5 = 1 \times 32 = 32$$
$$G_6 = ar^6 = 1 \times 64 = 64$$
$$G_7 = ar^7 = 1 \times 128 = 128$$

Hence, required GMs are: 2, 4, 8, 16, 32, 64, 128.

Example 1.32: Sum the series $1 + 3x + 5x^2 + 7x^3 + \cdots$ up to n terms, $x \neq 1$.

Solution: Note that nth term of this series $= (2n-1)\,x^{n-1}$.

Let $S_n = 1 + 3x + 5x^2 + \cdots + (2n-1)\,x^{n-1}$

Then, $xS_n = x + 3x^2 + \cdots + (2n-3)\,x^{n-1} + (2n-1)\,x^n$

Subtracing, we get:

$$S_n(1-x) = 1 + 2x + 2x^2 + \cdots + 2x^{n-1} + (2n-1)\,x^n$$

$$= 1 + 2x \times \frac{1-x^{n-1}}{1-x} - (2n-1)\,x^n$$

$$= \frac{1 - x + 2x - 2x^n - (2n-1)\,x^n(1-x)}{1-x}$$

$$= \frac{1 + x - 2x^n - (2n-1)\,x^n(2n-1)\,x^{n+1}}{1-x}$$

$$= \frac{1 + x - (2n+1)\,x^n + (2n-1)\,x^{n+1}}{1-x}$$

Hence
$$S = \frac{1 + x - (2n+1)\,x^n + (2n-1)\,x^{n+1}}{(1-x)^2}$$

Example 1.33: If in a G.P, $(p+q)$th term $= m$ and $(p-q)$th term $= n$, then find its pth and qth terms.

Solution: Suppose that the given GP be $a, ar, ar^2, ar^3, \ldots$

By hypothesis, $(p+q)$th term $= m = ar^{p+q-1}$

$$(p-q)\text{th term} = n = ar^{p-q-1}$$

Then, $\frac{m}{n} = r^{2q} \Rightarrow r = \left(\frac{m}{n}\right)^{1/2q}$

Hence, $m = a\left(\frac{m}{n}\right)^{(p+q-1)/2q} \Rightarrow a = m^{(q-p+1)/2q}\, n^{(p+q-1)/2q}$.

NOTES

Thus, pth term $= ar^{p-1} = m^{1/2}\, n^{1/2}$

$= \sqrt{mn}$

qth term $= ar^{q-1} = m^{\frac{2q-p}{2p} - n\frac{p}{2q}}$

Example 1.34: Sum the series $5 + 55 + 555 + \cdots$ up to n terms.

Solution: Let $S_n = 5 + 55 + 555 + \cdots$

$$S_n = 5\,(1 + 11 + 111 + \cdots)$$

$$= \frac{5}{9}\,(9 + 99 + 999 + \cdots)$$

$$= \frac{5}{9}\,[(10 - 1) + (100 - 1) + (1000 - 1) + \cdots]$$

$$= \frac{5}{9}\,[(10 + 10^2 + 10^3 + \cdots + 10^n) - (1 + 1 + \cdots \text{ upto } n \text{ terms})]$$

$$= \frac{5}{9}\,[(10 + 10^2 + 10^3 + \cdots + 10^n) - n]$$

$$= \frac{5}{9}\left[\frac{10(1-10^n)}{1-10} - n\right]$$

$$= \frac{5}{9}\left[\frac{10(10^n - 1)}{9} - n\right]$$

$$= \frac{50}{81}(10^n - 1) - \frac{5n}{9}$$

Example 1.35: If a, b, c, d are in GP, prove that $a^2 - b^2$, $b^2 - c^2$ and $c^2 - d^2$ are also in GP.

Solution: Since $\frac{b}{a} = \frac{c}{b} = \frac{d}{c} = k$ (say)

We have: $b = ak$, $c = bk$, $d = ck$

i.e. $b = ak$, $c = ak^2$, $d = ak^3$

Now, $(b^2 - c^2)^2 = (a^2k^2 - a^2k^4)^2$

$= a^4k^4(1 - k^2)^2$

Also, $(a^2 - b^2)\,(c^2 - d^2) = (a^2 - a^2k^2)\,(a^2k^4 - a^2k^6)$

$= a^4(1 - k^2)\,(k^4 - k^6)$

$= a^4k^2\,(1 - k^2)^2$

Hence, $(b^2 - c^2) = (a^2 - b^2)\,(c^2 - d^2)$.

This gives that $a^2 - b^2$, $b^2 - c^2$ and $c^2 - d^2$ are in GP

Example 1.36: Three numbers are in GP. Their product is 64 and sum is $\frac{124}{5}$. Find the numbers.

NOTES

Solution: Let the numbers be $\frac{a}{r}$, a, and ar.

Since $\frac{a}{r} + a + a^2 = \frac{124}{5}$ and $\frac{a}{r} \cdot a \cdot ar = 64$

We have $a^3 = 64 \Rightarrow a = 4$

This gives: $\frac{4}{r} + 4 + 4r = \frac{124}{5}$

$$\Rightarrow \quad \frac{1}{r} + 1 + r = \frac{31}{5}$$

$$\Rightarrow \quad \frac{r^2 + 1}{r} = \frac{26}{5}$$

$$\Rightarrow \quad 5r^2 + 5 = 26r$$

$$\Rightarrow \quad r = \frac{1}{5}$$

In either case, the numbers are $\frac{4}{5}$, 4, and 20.

Example 1.37: If a, b, c are in GP and $a^x = b^y = c^z$, prove that

$$\frac{1}{x} + \frac{1}{z} = \frac{2}{y}$$

Solution: a, b, c are in GP, $b^2 = ac$

But $\quad b^y = a^x \quad \Rightarrow \quad a = b^{y/x}$

and $\quad b^y = c^z \quad \Rightarrow \quad c + b^{y/z}$

So, we get, $\quad b^z = b^{y/x} \cdot b^{y/z}$

$$= b^{y\left(\frac{1}{x} + \frac{1}{z}\right)}$$

$$\Rightarrow \quad 2 = y\left(\frac{1}{x} + \frac{1}{z}\right)$$

$$\Rightarrow \quad \frac{1}{x} + \frac{1}{z} = \frac{2}{y}$$

Example 1.38: Sum to n terms the series

$$.7 + .77 + .777 + \cdots$$

Solution: Given series

$$= .7 + .77 + .777 + \cdots \text{ up to } n \text{ terms}$$

$$= 7\,(.1 + .11 + .111 + \cdots \text{ up to } n \text{ terms})$$

$$= \frac{7}{9}\,(.9 + .99 + .999 + \cdots \text{ up to } n \text{ terms})$$

$$= \frac{7}{9}\left[\left(1 - \frac{1}{10}\right) + \left(1 - \frac{1}{10^2}\right) + \left(1 - \frac{1}{10^3}\right) + \cdots\right]$$

NOTES

$$= \frac{7}{9}\left[n - \frac{1}{10} + \frac{1}{10^2} + \text{... up to } n \text{ terms} \right]$$

$$= \frac{7}{9}\left[\frac{\frac{1}{10}(1 - 1/10^n)}{1 - \frac{1}{10}} \right]$$

$$= \frac{7}{9}\left[n - \frac{1}{9}\left(1 - \frac{1}{10^n}\right) \right]$$

$$= \frac{7}{9}\left[n - \frac{1}{9}\left(1 - \frac{1}{10^n}\right) \right]$$

Example 1.39: The sum of three numbers in GP is 35 and their product is 1000. Find the numbers.

Solution: Let the numbers be $\frac{\alpha}{r}$, α, αr

Their product $\alpha^3 = 1000$

$\Rightarrow$ $\alpha = 10$

So the numbers are $\frac{10}{r}$, 10, $10r$

The sum of these numbers = 35

$\Rightarrow$ $\frac{10}{r} + 10 + 10r = 35$

$\Rightarrow$ $\frac{2}{r} + 2r = 5$

$\Rightarrow$ $2r^2 - 5r + 2 = 0$

$\Rightarrow$ $(2r - 1)(r - 2) = 0$

$\Rightarrow$ $r = 2$ or $\frac{1}{2}$

$r = 2$ gives the numbers as 5, 10, 20

$r = \frac{1}{2}$ gives the numbers as 20, 10, 5, the same as the first set.

Hence, the required numbers are 5, 10 and 20.

Example 1.40: The sum of the first eight terms of a GP (of real terms) is five time the sum of the first four terms. Find the common ratio.

Solution: Let the GP be a, ar, ar^2, . . .

$$S_8 = \text{Sum of first eight terms} = \frac{a(1 - r^8)}{1 - r}$$

$$S_4 = \text{Sum of first four terms} = \frac{a(1 - r^4)}{1 - r}$$

By hypothesis $S_8 = 5S_4 \Rightarrow \frac{a(1-r^8)}{1-r} = \frac{5a(1-r^4)}{1-r}$

$\Rightarrow \quad 1 - r^8 = 5(1 - r^4)$

$\Rightarrow \quad (1 - r^4)(1 + r^4) = 5(1 - r^4)$

In case $r^4 - 1 = 0$, we get $(r^2 - 1) = 0 \Rightarrow r = \pm 1$

(Note that $r^2 + 1 = 0 \Rightarrow r$ is imaginary)

Now, $r = 1 \Rightarrow$ the given series is $a + a + a + \cdots$

but then $S_8 = 8a$ and $S_4 = 4a$.

So, $S_8 \neq 4S_4$

In case $r = -1$, we get $S_8 = 0$ and $S_4 = 0$

Hence, the hypothesis is satisfied.

Now, suppose $r^4 - 1 \neq 0$, then $1 + r^4 = 5$

$\Rightarrow \quad r^4 = 4 \Rightarrow r^2 = 2 \quad (r^2 \neq -2)$

$\Rightarrow r = \pm\sqrt{2}$

Hence, $r = -1$ or $\pm\sqrt{2}$

Example 1.41: If S is the sum, P the product and R the sum of reciprocals of n terms in GP, prove that

$$P^2R^n = S^n$$

Solution: Let $a, ar, ar^2, \cdots$ be the given GP.

Then $S = a + ar + ar^2 + \cdots$ up to n terms

$$= \frac{a(1-r^n)}{1-r} \qquad \text{(i)}$$

$$p = a \cdot ar \cdot ar^2 \cdots ar^{n-1}$$

$$= a^n \, r^{1+2+3+\cdots+(n-1)}$$

$$= a^n \, r^{\frac{(n-1)}{2}(1+n-1)}$$

$$= a^n \, r^{\left(\frac{n-1}{2}\right)^n} \qquad \text{(ii)}$$

$$R = \frac{1}{a} + \frac{1}{ar} + \frac{1}{ar^2} + \cdots \quad \text{up to } n \text{ terms}$$

$$= \frac{\frac{1}{a}\left(1 - \frac{1}{r^n}\right)}{1 - \frac{1}{r}} = \frac{r}{a}\frac{(r^n - 1)}{(r-1)\,r^n}$$

$$= \frac{(1-r^n)}{a(1-r)\,r^{n-1}} \qquad \text{(iii)}$$

NOTES

NOTES

By (ii) and (iii), $P^2R^n = a^{2n}\ r^{n(n-1)}\ \dfrac{(1-r^n)^n}{a^n\,(1-r)^n\ r^{n(n-1)}}$

$$= \frac{a^n(1-r^n)^n}{(1-r)^n} = S^n \text{ by (i)}$$

Example 1.42: The ratio of the 4th to the 12th term of a GP with positive common ratio is $\frac{1}{256}$. If the sum of the two terms is 61.68, find the sum of series to 8 terms.

Solution: Let the series be $a, ar, ar^2, \ldots,$

$$T_4 = 4\text{th term} = ar^3$$

$$T_{12} = 12\text{th term} = ar^{11}$$

By hypothesis $\dfrac{T_4}{T_{12}} = \dfrac{1}{256}$

i.e. $\dfrac{ar^3}{ar^{11}} = \dfrac{1}{256}$

$$\frac{1}{r^8} = \frac{1}{256}$$

$\Rightarrow$ $r^8 = 256$

$\Rightarrow$ $r = \pm 2$

Since r is given to be positive, we reject negative sign.

Again it is given that

$$T_4 + T_{12} = 61.68$$

i.e. $a\,(r^3 + r^{11}) = 61.68$

$$a\,(8 + 2048) = 61.68$$

$$a = \frac{61.68}{2056} = .03$$

Hence, $S_8 =$ sum to eight terms

$$= \frac{a\,(1-r^8)}{1-r} = \frac{a\,(r^8-1)}{r-1}$$

$$= \frac{(.03)\,(256-1)}{(2-1)}$$

$$= .03 \times 255$$

$$= 7.65$$

Example 1.43: A manufacturer reckons that the value of a machine which costs him Rs 18,750 will depreciate each year by 20%. Find the estimated value at the end of 5 years.

NOTES

Solution: At the end of first year the value of machine

$$= 18{,}750 \times \frac{80}{100}$$

$$= \frac{4}{5}(18{,}750)$$

At the end of second year, it is equal to $\left(\frac{4}{5}\right)^2 (18{,}750)$

Proceeding in this manner, the estimated value of machine at the end of 5 years is $\left(\frac{4}{5}\right)^5 (18{,}750)$

$$= \frac{64 \times 16}{125 \times 25} \times 18{,}750$$

$$= \frac{1024}{125} \times 750$$

$$= (1024) \times 6$$

$$= \text{Rs } 6144$$

Example 1.44: Show that a given sum of money accumulated at 20 per cent per annum more than doubles itself in 4 years at compound interest.

Solution: Let the given sum be a rupees. After 1 year, it becomes $\frac{6a}{5}$ (it is increased by $\frac{a}{5}$).

At the end of two years, it becomes $\frac{6}{5}\left(\frac{6a}{5}\right) = \left(\frac{6}{5}\right)^2 a$

Proceeding in the manner, at the end of 4th year, the amount will be $\left(\frac{6}{5}\right)^4 a = \frac{1296}{625} a$.

Now, $\frac{1296}{625} a - 2a = \frac{46}{625} a,$ a positive quantity, so the amount after 4 years is more than double of the original amount.

Example 1.45: If $x = a + \frac{a}{r} + \frac{a}{r^2} + \cdots \infty$

$$y = b - \frac{b}{r} + \frac{b}{r^2} + \cdots \infty$$

and $z = c + \frac{c}{r^2} + \frac{c}{r^4} + \cdots \infty$

Show that $\frac{xy}{z} = \frac{ab}{c}$

NOTES

Solution: Clearly, $x = \dfrac{a}{1-\dfrac{1}{r}} = \dfrac{ar}{r-1}$

$$y = \frac{b}{1-(-1/r)} = \frac{br}{r+1}$$

and $$z = \frac{c}{1-\dfrac{1}{r^2}} = \frac{cr^2}{r^2-1}$$

Now, $$\frac{xy}{z} = \frac{ab\,r^2}{(r^2-1)} \Big/ \left(\frac{cr^2}{r^2-1}\right)$$

$$= \frac{ab}{c}$$

Example 1.46: If $a^2 + b^2$, $ab + bc$ and $b^2 + c^2$ are in GP, prove that a, b, c are also in GP.

Solution: Since $a^2 + b^2$, $ab + bc$ and $b^2 + c^2$ are in GP, we get:

$$(ab + bc)^2 = (a^2 + b^2)(b^2 + c^2)$$

$$b^2(a^2 + 2ac + c^2) = a^2b^2 + a^2c^2 + b^4 + b^2c^2$$

$$\Rightarrow \quad 2ab^2c^2 = a^2c^2 + b^4$$

$$\Rightarrow a^2c^2 - 2ab^2c^2 + b^4 = 0$$

$$\Rightarrow \quad (ac - b^2)^2 = 0$$

$$\Rightarrow \quad ac = b^2$$

$$\Rightarrow \quad a, b, c \text{ are in GP}$$

CHECK YOUR PROGRESS

7. What is geometric progression?
8. Define geometric mean.

1.5 MATRICES

1.5.1 What is a Matrix?

A matrix (plural matrices) is a rectangular array of numbers. Let F be a field and n, m be two integers ≥ 1. An array of elements in F, of the type

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

is called a matrix in F. We denote this matrix by (a_{ij}), $i = 1, ..., m$ and $j = 1, ..., n$. We say that it is an $m \times n$ matrix (or matrix of order $m \times n$). It has m rows and n columns. For example, the first row is a_{11}, $(a_{12}, ..., a_{1n})$ and first column is

$$\begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix}$$

Also, a_{ij} denotes the element of the matrix (a_{ij}) lying in ith row and jth column and we call this element as the (i, j)th element of the matrix.

For example, in the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

$a_{11} = 1$, $a_{12} = 2$, $a_{32} = 8$, i.e. $(1, 1)$th element is 1

$(1, 2)$th element is 2.

$(3, 2)$th element is 8.

Notes: (i) Unless otherwise stated, you should consider matrices over the field C of complex numbers only.

(Note the field of real numbers is contained in C.)

(ii) A matrix is simply an arrangement of elements and has no numerical value.

1.5.2 Types of Matrices

The various types of matrices are as follows:

1. **Row matrix:** A matrix which has exactly one row is called a row matrix.

   For example, (1, 2, 3, 4) is a row matrix.

2. **Column matrix:** A matrix which has exactly one column is called a column matrix.

   For example, $\begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix}$ is a column matrix.

3. **Square matrix:** A matrix in which the number of rows is equal to the number of columns is called a square matrix.

   For example, $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ is a 2×2 square matrix.

4. **Null or zero matrix:** A matrix each of whose elements is zero is called a null matrix or zero matrix.

   For example, $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ is a 2×3 null matrix.

NOTES

5. **Diagonal matrix:** The elements a_{ij} are called diagonal elements of a square matrix (a_{ij}). For example, in

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

the diagonal elements are $a_{11} = 1$, $a_{22} = 5$, $a_{33} = 9$

A square matrix whose every element other than diagonal elements is zero, is called a diagonal matrix. For example,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$ is a diagonal matrix.

Note that the diagonal elements in a diagonal matrix may also be zero. For example

$\begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ are also diagonal matrices.

6. **Scalar matrix:** A diagonal matrix whose diagonal elements are equal is called a scalar matrix. For example,

$$\begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$ are scalar matrices.

7. **Identity matrix:** A diagonal matrix whose diagonal elements are all equal to 1 (unity) is called identity matrix or unit matrix. For example,

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$ is an identity matrix.

8. **Triangular matrix:** A square matrix (a_{ij}) whose elements $a_{ij} = 0$ wherever $i < j$ is called a lower triangular matrix.

Similarly, a square matrix (a_{ij}) whose elements $a_{ij} = 0$ whenever $i > j$ is called an upper triangular matrix.

For example,

$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 5 & 0 \\ 7 & 8 & 9 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}$ are lower triangular matrices

and $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$ are upper triangular matrices.

1.5.3 Algebra of Matrices

Equality of matrices

Two matrices A and B are said to be equal if

(*i*) A and B are of same order.

NOTES

(*ii*) Corresponding elements in A and B are same. For example, the following two matrices are equal:

$$\begin{pmatrix} 3 & 4 & 9 \\ 16 & 25 & 64 \end{pmatrix} = \begin{pmatrix} 3 & 4 & 9 \\ 16 & 25 & 64 \end{pmatrix}$$

But the following two matrices are not equal:

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

as matrix on left is of order 2×3, while on right it is of order 3×3.

The following two matrices are also not equal:

$$\begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 4 & 8 & 9 \end{pmatrix}$$

as (2, 1)th element in LHS matrix is 7, while in RHS matrix it is 4.

Addition of matrices

If A and B are two matrices of the same order, then addtion of A and B is defined to be the matrix obtained by adding the corresponding elements of A and B.

For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, B = \begin{pmatrix} 2 & 3 & 4 \\ 5 & 6 & 7 \end{pmatrix}$$

Then $A + B = \begin{pmatrix} 1+2 & 2+3 & 3+4 \\ 4+5 & 5+6 & 6+7 \end{pmatrix} = \begin{pmatrix} 3 & 5 & 7 \\ 9 & 11 & 3 \end{pmatrix}$

Also $A - B = \begin{pmatrix} 1-2 & 2-3 & 3-4 \\ 4-5 & 5-6 & 6-7 \end{pmatrix} = \begin{pmatrix} -1 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix}$

Note that addition (or subtraction) of two matrices is defined only when A and B are of the same order.

Properties of matrix addition

(*i*) Matrix addition is commutative.

i.e. $A + B = B + A$

For, (i, j)th element of $A + B$ is $(a_{ij} + b_{ij})$ and of $B + A$ is $(b_{ij} + a_{ij})$, and they are same as a_{ij}, and b_{ij} are complex numbers.

(*ii*) Matrix addition is associative.

i.e. $A + (B + C) = (A + B) + C$

For, (i, j)th element of $A + (B + C)$ is $a_{ij} + (b_{ij} + c_{ij})$ and of $(A + B) + C$ is $(a_{ij} + b_{ij}) + c_{ij}$, which are the same.

(*iii*) If O denotes null matrix of the same order as that of A, then

$A + O = A = O + A$

For, (i, j)th element of $A + O$ is $a_{ij} + O + a_{ij}$, which is same as (i, j)th element of A.

NOTES

(*iv*) To each matrix A, there corresponds a matrix B such that $A + B = O = B + A$.

For, let (i, j)th element of B be $-a_{ij}$. Then (i, j)th element of A + B is $a_{ij} - a_{ij} = 0$.

Thus, the set of $m \times n$ matrices forms an Abelian group under the composition of matrix addition.

Multiplication of matrix by scalar

If k is any complex number and A a given matrix, then kA is the matrix obtained from A by multiplying each element of A by k. The number k is called *scalar*.

For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \text{ and } k = 2$$

then $$kA = \begin{pmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{pmatrix}$$

It can be easily shown:

(*i*) $k(A + B) = kA + kB$ (*ii*) $(k_1 + k_2)A = k_1A + k_2A$

(*iii*) $1A = A$ (*iv*) $(k_1k_2)A = k_1(k_2A)$.

Example 1.47: If $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$ and $B = \begin{pmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \end{pmatrix}$,

verify $A + B = B + A$.

Solution: $$A + B = \begin{pmatrix} 1+0 & 2+1 & 3+2 \\ 4+3 & 5+4 & 6+5 \end{pmatrix} = \begin{pmatrix} 1 & 3 & 5 \\ 7 & 9 & 11 \end{pmatrix}$$

$$B + A = \begin{pmatrix} 0+1 & 1+2 & 2+3 \\ 3+4 & 4+5 & 5+6 \end{pmatrix} = \begin{pmatrix} 1 & 3 & 5 \\ 7 & 9 & 11 \end{pmatrix}$$

So, $A + B = B + A$

Example 1.48: If A and B are matrices as in Example 1.47

and $C = \begin{pmatrix} -1 & 0 & 1 \\ 1 & 2 & 3 \end{pmatrix}$, verify $(A + B) + C = A + (B + C)$

Solution: Now $A + B = \begin{pmatrix} 1 & 3 & 5 \\ 7 & 9 & 11 \end{pmatrix}$

So, $$(A + B) + C = \begin{pmatrix} 1-1 & 3+0 & 5+1 \\ 7+1 & 9+2 & 11+3 \end{pmatrix} = \begin{pmatrix} 0 & 3 & 6 \\ 8 & 11 & 14 \end{pmatrix}$$

Again, $$B + C = \begin{pmatrix} 0-1 & 1+0 & 2+1 \\ 3+1 & 4+2 & 5+3 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 3 \\ 4 & 6 & 8 \end{pmatrix}$$

So, $$A + (B + C) = \begin{pmatrix} 1-1 & 2+1 & 3+3 \\ 4+4 & 5+6 & 6+8 \end{pmatrix} = \begin{pmatrix} 0 & 3 & 6 \\ 8 & 11 & 14 \end{pmatrix}$$

Therefore, $(A + B) + C = A + (B + C)$

NOTES

Example 1.49 If $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$, find a matrix B such that $A + B = 0$.

Solution: Let $B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}$

Then $A + B = \begin{pmatrix} 1+b_{11} & 2+b_{12} \\ 3+b_{21} & 4+b_{22} \\ 5+b_{31} & 6+b_{32} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$

implies, $b_{11} = -1$, $b_{12} = -2$, $b_{21} = 3$, $b_{22} = -4$ and $b_{31} = -5$, $b_{32} = -6$.

Therefore, required $B = \begin{pmatrix} -1 & -2 \\ -3 & -4 \\ -5 & -6 \end{pmatrix}$.

Example 1.50: (*i*) If $A = \begin{pmatrix} 0 & 1 & 2 \\ 2 & 3 & 4 \\ 4 & 5 & 6 \end{pmatrix}$ and $k_1 = i$, $k_2 = 2$, verify $(k_1 + k_2)$ $A = k_1A + k_2A$.

(*ii*) If $A = \begin{pmatrix} 0 & 2 & 3 \\ 2 & 1 & 4 \end{pmatrix}$, $B = \begin{pmatrix} 7 & 6 & 3 \\ 1 & 4 & 5 \end{pmatrix}$ find the value of $2A + 3B$.

Solution: (*i*) $k_1A = \begin{pmatrix} 0 & i & 2i \\ 2i & 3i & 4i \\ 4i & 5i & 6i \end{pmatrix}$ and $k_2A = \begin{pmatrix} 0 & 2 & 4 \\ 4 & 6 & 8 \\ 8 & 10 & 12 \end{pmatrix}$

So, $k_1A + k_2A = \begin{pmatrix} 0 & 2+i & 4+2i \\ 4+2i & 6+3i & 8+4i \\ 8+4i & 10+5i & 12+6i \end{pmatrix}$

Also, $(k_1 + k_2)\,A = \begin{pmatrix} 0 & 2+i & 4+2i \\ 4+2i & 6+3i & 8+4i \\ 8+4i & 10+5i & 12+6i \end{pmatrix}$

Therefore, $(k_1 + k_2)A = k_1A + k_2A$

(*ii*) $2A = \begin{pmatrix} 0 & 4 & 6 \\ 4 & 1 & 8 \end{pmatrix}$

$3B = \begin{pmatrix} 21 & 18 & 9 \\ 3 & 12 & 15 \end{pmatrix}$

So, $2A + 3B = \begin{pmatrix} 21 & 22 & 15 \\ 7 & 13 & 23 \end{pmatrix}$.

NOTES

Example 1.51: If $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 0 & 1 & 2 \end{pmatrix}$ find $a_{11}, a_{22}, a_{33}, a_{31}, a_{41}$.

Solution: a_{11} = element of A in first row and first column = 1

a_{22} = element of A in second row and second column = 5

a_{33} = element of A in third row and thrid column = 9

a_{31} = element of A in third row and first column = 7

a_{41} = element of A in fourth row and first column = 0

Multiplication of matrices

The product AB of two matrices, A and B, is defined only when the number of columns of A is same as the number of rows in B and by definition the product AB is a matrix C of order $m \times p$ if A and B were of order $m \times n$ and $n \times p$ respectively. The following example illustrates the rule to multiply two matrices:

Let, $A = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{pmatrix}$, $B = \begin{pmatrix} d_1 & e_1 \\ d_2 & e_2 \\ d_3 & e_3 \end{pmatrix}$

Order of $A = 2 \times 3$ and Order of $B = 3 \times 2$

So, AB is defined as:

$$G = AB = \begin{pmatrix} a_1d_1 + b_1d_2 + c_1d_3 & a_1e_1 + b_1e_2 + c_1e_3 \\ a_2d_1 + b_2d_2 + c_2d_3 & a_2e_1 + b_2e_2 + c_2e_3 \end{pmatrix}$$

$$= \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}$$

where to get

g_{11} : Multiply elements of the first row of A with corresponding elements of the first column of B and add.

g_{12} : Multiply elements of the first row of A with corresponding elements of the second column of B and add.

g_{21} : Multiply elements of the second row of A with corresponding elements of the first column of B and add.

g_{22} : Multiply elements of the second columns of A with corresponding elements of the second column and add.

Notes:

(*i*) In general, if A and B are two matrices, then AB may not be equal to BA. For example, if

$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, then $AB = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$

and $BA = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$. So, $AB \neq BA$

(*ii*) If product *AB* is defined, then it is not necessary that *BA* must also be defined. For example, if *A* is of order 2 × 3 and *B* is of order 3 × 1, then clearly *AB* is defined but *BA* is not defined (as the number of columns of *B* ≠ the number of rows of *A*).

It can easily be verified that (whenever products are defined).

$A(BC) = (AB)C$

$A(B + C) = AB + AC$

$(A + B)C = AC + BC$

Example 1.52: If $A = \begin{pmatrix} 2 & -1 \\ 0 & 3 \end{pmatrix}$ and $B = \begin{pmatrix} 7 & 0 \\ -2 & -3 \end{pmatrix}$ write down *AB*.

Solution: $AB = \begin{pmatrix} 2 \times 7 + (-1) \times (-2) & 2 \times 0 + (-1) \times (-3) \\ 0 \times 7 + 3 \times (-2) & 0 \times 0 + 3 \times (-3) \end{pmatrix}$

$= \begin{pmatrix} 16 & 3 \\ -6 & 9 \end{pmatrix}$

Example 1.53: Verify the associative law $A(BC) = (AB)C$ for the following matrices.

$$A = \begin{pmatrix} -1 & 0 & 5 \\ 7 & -2 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 0 & 5 \\ 7 & -2 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} -1 & -1 \\ 2 & 0 \\ 0 & 4 \end{pmatrix}$$

Solution: $AB = \begin{pmatrix} 4 & -7 & 25 \\ 13 & 51 & 0 \end{pmatrix}$

So, $(AB)C = \begin{pmatrix} -18 & 96 \\ 89 & -13 \end{pmatrix}$

Again, $BC = \begin{pmatrix} 13 & -1 \\ 1 & 3 \\ -1 & 19 \end{pmatrix}$

So, $A(BC) = \begin{pmatrix} -18 & 96 \\ 89 & -13 \end{pmatrix}$

Therefore, $A(BC) = (AB)C$

Example 1.54: If *A* is a square matrix, then *A* can be multiplied by itself. Define $A^2 = A.A$ (called power of a matrix). Compute A^2 for the following matrix:

$$A = \begin{pmatrix} 1 & 0 \\ 3 & 4 \end{pmatrix}$$

Solution: $A^2 = \begin{pmatrix} 1 & 0 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 15 & 16 \end{pmatrix}$

(Similarly, we can define $A^3, A^4, A^5, \ldots$ for any square matrix *A*.)

NOTES

Example 1.55: If $A = \begin{pmatrix} 1 & 2 \\ -3 & 0 \end{pmatrix}$ find $A^2 + 3A + 5I$ where I is unit matrix of order 2.

Solution: $A^2 = \begin{pmatrix} 1 & 2 \\ -3 & 0 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ -3 & 0 \end{pmatrix} = \begin{pmatrix} -5 & 2 \\ -3 & -6 \end{pmatrix}$

$$3A = \begin{pmatrix} 3 & 6 \\ -9 & 0 \end{pmatrix}$$

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

So, $A^2 + 3A + I = \begin{pmatrix} 5 & 2 \\ -3 & -6 \end{pmatrix} + \begin{pmatrix} 3 & 7 \\ -9 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$$= \begin{pmatrix} -1 & 8 \\ -12 & -5 \end{pmatrix}$$

Example 1.56: If $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$

show that $AB = -BA$ and $A^2 = B^2 = I$.

Solution: $AB = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$

$$BA = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}$$

So, $AB = -BA$

Also, $A^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$

$$B^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$$

This proves the result.

1.5.4 Transpose of a Matrix

Let A be a matrix. The matrix obtained from A by interchange of its rows and column, is called the *transpose* of A. For example,

if $A = \begin{pmatrix} 1 & 0 & 2 \\ 2 & 1 & 0 \end{pmatrix}$, then transpose of $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 2 & 0 \end{pmatrix}$

Transpose of A is denoted by A'.

It can be easily verified that:

(*i*) $(A')' = A$

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

(*iii*) $(AB)' = B'A'$

NOTES

Example 1.57: For the following matrices A and B verify $(A + B) = A' + B'$.

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 3 & 4 \\ 1 & 8 & 6 \end{pmatrix}$$

Solution: $A' = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$ $B' = \begin{pmatrix} 2 & 1 \\ 3 & 8 \\ 4 & 6 \end{pmatrix}$

So, $A' + B' = \begin{pmatrix} 3 & 5 \\ 5 & 13 \\ 7 & 12 \end{pmatrix}$

Again, $A + B = \begin{pmatrix} 3 & 5 & 7 \\ 5 & 13 & 12 \end{pmatrix}$

So, $(A + B)' = \begin{pmatrix} 3 & 5 \\ 5 & 13 \\ 7 & 12 \end{pmatrix}$

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

1.5.5 Elementary Operations

Consider the following matrices:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad B = \begin{pmatrix} 4 & 5 & 6 \\ 1 & 2 & 3 \end{pmatrix}$$

Matrix B is obtained from A by interchange of first and second rows.

Consider $C = \begin{pmatrix} 2 & 1 & 3 \\ 3 & 4 & 5 \end{pmatrix}$ and $D = \begin{pmatrix} 6 & 3 & 9 \\ 3 & 4 & 5 \end{pmatrix}$

Matrix D is obtained from C by multiplying the first row by 3.

Consider $E = \begin{pmatrix} 2 & 3 & 4 \\ 1 & 3 & 2 \end{pmatrix}$ and $F = \begin{pmatrix} 2 & 3 & 4 \\ 7 & 12 & 14 \end{pmatrix}$

Matrix F is obtained from E by multiplying the first row of E by 3 and adding it to second row.

Such operations on rows of a matrix as described above are called *elementary row operations.*

Similarly, we define *elementary column operations.*

An elementary operation is either elementary row operation or elementary column operation and is of the following three types:

Type I. The interchange of any two rows (or column).

Type II. The multiplication of any row (or column) by a non-zero number.

Type III. The addition of multiple of one row (or column) to another row (or column).

We shall use the following notations for three types of elementary operations.

(*i*) The interchange of *i*th and *j*th rows (columns) will be denoted by $R_i \leftrightarrow R_j$ ($C_i \leftrightarrow C_j$).

NOTES

(*ii*) The multiplication of *i*th row (column) by non-zero number *k* will be denoted by $R_i \to k\,R_i$ $(C_i \to k\,C_i)$

(*iii*) The addition of *k* times the *j*th row (column) to *i*th raw (column) will be denoted by $R_i \to R_i + kR_j$ $(C_i \to C_i + kC_i)$.

1.5.6 Elementary Matrices

Matrix obtained from an identity matrix by a single elementary operation is called **elementary matrix**.

For example, $\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

are elementary matrices, the first is obtained by $R_1 \leftrightarrow R_2$ and the second by $C_1 \to 2C_1$ on the identity matrix.

We state the following result without proof:

'An elementary row operation on the product of two matrices is equivalent to the elementary row operation on prefactor.'

It means that if we make elementary row operation in the product *AB*, then it is equivalent to making the same elementary row operation in *A* and then multiplying it with *B*.

Let $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 2 & 3 \end{pmatrix}$

Then $AB = \begin{pmatrix} 9 & 13 \\ 13 & 19 \end{pmatrix}$

Suppose we interchange first and second rows of *AB*.

Then, the matrix we get is:

$$C = \begin{pmatrix} 13 & 19 \\ 9 & 13 \end{pmatrix}$$

Now interchange the first and second rows of *A* and get a new matrix

$$D = \begin{pmatrix} 2 & 3 & 4 \\ 1 & 2 & 3 \end{pmatrix}$$

Multiply *D* with *B* to get

$$DB = \begin{pmatrix} 2 & 3 & 4 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 13 & 19 \\ 9 & 13 \end{pmatrix}$$

So, $DB = C$

1.5.7 Gauss Elimination Method

Suppose we have a system of equations in the matrix form $AX = B$, where

$$A = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix}, \quad X = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \text{ and } B = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}$$

NOTES

The matrix

$$C = [A/B] = \left(\begin{array}{ccc|c} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \end{array}\right)$$

is called the augmented matrix of the given system of equations. Instead of writing the whole equation $AX = B$ and making elementary row transformations to it, we sometimes work only on the augmented matrix and apply these operations on this matrix to get the solution. We explain this method by considering the following example.

Suppose we have the system of equations

$$x + y + z = 7$$
$$x + 2y + 3z = 16$$
$$x + 3y + 4z = 22$$

Which in the matrix form will be $AX = B$, where

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \end{pmatrix}, \quad X = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \text{ and } B = \begin{pmatrix} 7 \\ 16 \\ 22 \end{pmatrix}$$

Augmented matrix of this system of equations is:

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 7 \\ 1 & 2 & 3 & 16 \\ 1 & 3 & 4 & 22 \end{array}\right)$$

Which becomes:

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 7 \\ 0 & 1 & 2 & 9 \\ 1 & 2 & 3 & 15 \end{array}\right) \qquad [R_2 \to R_2 - R_1, R_3 \to R_3 - R_1]$$

or

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 7 \\ 0 & 1 & 2 & 9 \\ 0 & 0 & -1 & -3 \end{array}\right) \qquad [R_3 \to R_3 - 2R_1]$$

Hence, we get $Z = 3$, $y + 2z = 9$, $x + y + z = 7$

This gives us the solution $x = 1$, $y = 3$ and $z = 3$.

Thus, it requires the same operations as were used earlier. It is only a different way of expressing the same thing.

This method is called the Gauss elimination method of solving equations.

Sometimes, we proceed further and reduce the augmented matrix to

$$\left(\begin{array}{ccc|c} 1 & 0 & -1 & -2 \\ 0 & 1 & 2 & 9 \\ 0 & 0 & -1 & -3 \end{array}\right) \qquad [R_1 \to R_1 - R_2]$$

NOTES

$$\left(\begin{array}{ccc|c} 1 & 0 & -1 & -2 \\ 0 & 1 & 3 & 12 \\ 0 & 0 & -1 & -3 \end{array}\right) \qquad [R_2 \to R_2 - R_3]$$

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 3 & 12 \\ 0 & 0 & -1 & -3 \end{array}\right)$$

or

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 3 \end{array}\right)$$

and the solution is $x = 1$, $y = 3$, $z = 3$

This method is called **Gauss Jordan Reduction**.

Note: A matrix is said to be in row echelon form if:

1. All rows in the matrix which consist of zeros are at the bottom of the matrix (such rows may or may not be there).
2. The first non-zero entry in each non-zero row is called the leading entry.
3. If kth and $(k + 1)$th rows are two consecutive rows (having some non-zero entry), then the leading entry of the $(k + 1)$th row is to the right of the leading entry of the kth row.

   Thus, the Gaussian elimination method requires the augmented matrix to be put in the row echelon form.

   In addition to the above three conditions, the matrix also satisfies the following:

4. If a column contains a leading entry of some row, then all other entries in that column are zero. Then we say that matrix is in reduced echelon form and this method is called Gauss-Jordan reduction.

We, thus, realize that the Gauss–Jordan reduction requires few extra steps than the Gauss elimination method. But then in the former case, the solution is obtained without any back substitution.

Example 1.58: The equilibrium conditions for two substitute goods are given by

$$5P_1 - 2P_2 = 15$$

$$-P_1 + 8P_2 = 16$$

Find the equilibrium prices.

Solution: We write the given system of equations in the matrix form:

$$\begin{pmatrix} 5 & -2 \\ -1 & 8 \end{pmatrix}\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \begin{pmatrix} 15 \\ 16 \end{pmatrix}$$

The augmented matrix is:

$$\left(\begin{array}{cc|c} 5 & -2 & 15 \\ -1 & 8 & 16 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & -2/5 & 3 \\ -1 & 8 & 6 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & -2/5 & 3 \\ 0 & \frac{38}{5} & 19 \end{array}\right)$$

$$R_1 \to \frac{R_1}{5} \qquad R_2 \to R_2 + R_1$$

Solution is then given by

$$\frac{38}{5}P_2 = 19 \text{ and } P_1 - \frac{2}{5}P_2 = 3$$

i.e. $$P_1 = 4 \text{ and } P_2 = \frac{5}{2}$$

Example 1.59: An automobile company uses three types of steel S_1, S_2 and S_3 for producing three types of cars C_1, C_2 and C_3. Steel requirement (in tons) for each type of car is given as

	C_1	C_2	C_3
S_1	2	3	4
S_2	1	1	2
S_3	3	2	1

Determine the number of cars of each type that can be producd using 29, 13 and 16 tons of steel of three types, respectively.

Solution: Suppose *x*, *y* and *z* are the number of cars of each type that are produced. Then,

$$\begin{aligned} 2x + 3y + 4z &= 29 \\ x + y + 2z &= 13 \\ 3x + 2y + z &= 16 \end{aligned}$$

This system of equations can be put in the matrix form as

$$\begin{pmatrix} 2 & 3 & 4 \\ 1 & 1 & 2 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 29 \\ 13 \\ 16 \end{pmatrix}$$

Augmented matrix is:

$$\left(\begin{array}{ccc|c} 2 & 3 & 4 & 29 \\ 1 & 1 & 2 & 13 \\ 3 & 2 & 1 & 16 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 1 & 2 & 13 \\ 2 & 3 & 4 & 29 \\ 3 & 2 & 1 & 16 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 1 & 2 & 13 \\ 0 & 1 & 0 & 3 \\ 0 & -1 & -5 & -23 \end{array}\right)$$

$$R_2 \leftrightarrow R_1 \qquad R_2 \to R_2 - 2R_1 \qquad R_3 \to R_3 - 3R_2$$

$$\sim \left(\begin{array}{ccc|c} 1 & 1 & 2 & 13 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & -5 & -20 \end{array}\right)$$

$$R_3 \to R_3 + R_2$$

We, thus, have $x + y + 2z = 13$

$y = 3$

$-5z = -20$

NOTES

Putting $z = 4$, $y = 3$ and $x = 2$, the required number of cars can be produced.

Example 1.60: A firm produces two products P_1 and P_2, passing through two machines M_1 and M_2 before completion. M_1 can produce either 10 units of P_1 or 15 units of P_2 per hour. M_2 can produce 15 units of either product per hour. Find daily production of P_1 and P_2 if the time available is 12 hours on M_1 and 10 hours on M_2 per day.

Solution: Suppose daily production of P_1 is x units and of P_2 is y units. Then

$$\frac{x}{10} + \frac{y}{15} = 12 \qquad \frac{x}{15} + \frac{y}{15} = 10$$

i.e. $3x + 2y = 360$

$x + y = 150$

Matrix representation is given by

$$\begin{pmatrix} 3 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 360 \\ 150 \end{pmatrix}$$

Augmented matrix is:

$$\left(\begin{array}{cc|c} 3 & 2 & 360 \\ 1 & 1 & 150 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & 1 & 150 \\ 3 & 2 & 360 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & 1 & 150 \\ 0 & -1 & -90 \end{array}\right)$$

$R_1 \leftrightarrow R_2 \qquad R_2 \rightarrow 3R_1$

i.e. $x + y = 150$

$-y = -90$ or that $x = 60$, $y = 90$ (daily production of P_1 and P_2)

Example 1.61: There are three types of foods, Food I, Food II, Food III. Food I contains I unit each of three nutrients *A, B, C*. Food II contains 1 unit of nutrient *A*, 2 units of nutrient *B* and 3 units of nutrient *C*. Food III contains 1, 3 and 4 units of nutrients *A, B, C*. 7 units of *A*, 16 units of *B* and 22 units of nutrient *C* are required. Find the amount of three foods that will provide these.

Solution: Suppose *x, y* and *z* are the amounts of three foods to be taken so as to get required nutrients.

Then $x + y + z = 7$

$x + 2y + 3z = 16$

$x + 3y + 4z = 22$

NOTES

In matrix form, we get

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 7 \\ 16 \\ 22 \end{pmatrix}$$

Augmented matrix is:

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 7 \\ 1 & 2 & 3 & 16 \\ 1 & 3 & 4 & 22 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 1 & 1 & 7 \\ 0 & 1 & 2 & 9 \\ 0 & 2 & 3 & 15 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 1 & 1 & 7 \\ 0 & 1 & 2 & 9 \\ 0 & 0 & -1 & -3 \end{array}\right)$$

$$R_2 \to R_2 - R_1 \qquad R_3 \to R_3 - 2R_1$$
$$R_3 \to R_3 - R_1$$

which gives $z = 3$, $y + 2z = 9$, $x + y + z = 7$

Hence, $x = 1$, $y = 3$, $z = 3$ is the required solution.

1.5.8 Reduction of a Matrix to Echelon Form

Consider $A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 2 \\ 3 & 1 & 2 & 4 \end{pmatrix}$

Apply the following elementary row operations on A

$$R_2 \to R_2 - 2R_1, \quad R_3 \to R_3 - 3R_1$$

and obtain a new matrix:

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

Apply $R_2 \to -\frac{1}{3} R_2$ on B to get

$$C = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 1 & 2 \\ 0 & -5 & -7 & -8 \end{pmatrix}$$

Apply $R_3 \to R_3 + 5R_2$ on C to get

$$D = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & -2 & 2 \end{pmatrix}$$

The matrix D is in *echelon form* (i.e. elements below the diagonal are zero).

We thus find elementary row operations reduce matrix A to echelon form.

In fact, *any matrix can be reduced to echelon form by elementary row operations.* The procedure is as follows:

Step I. Reduce the element in (1, 1)th place to unity by some suitable elementary row operation.

NOTES

Step II. Reduce all the elements in first column below first row to zero with the help of unity obtained in first step.

Step III. Reduce the element in (2, 2)th place to unity by suitable elementary row operations.

Step IV. Reduce all the elements in second column below second row to zero with the help of unity obtained in Step III.

Proceeding in this way, any matrix can be reduced to echelon form.

Example 1.62: Reduce $A = \begin{pmatrix} 3 & -10 & 5 \\ -1 & 12 & -2 \\ 1 & -5 & 2 \end{pmatrix}$ to echelon form.

Solution:

Step I. Apply $R_1 \leftrightarrow R_3$ to get

$$\begin{pmatrix} 1 & -5 & 2 \\ -1 & 12 & -2 \\ 3 & -10 & 5 \end{pmatrix}$$

Step II. Apply $R_2 \to R_2 + R_1, R_3 \to R_3 - 3R_1$ to get

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

Step III. Apply $R_2 \to \frac{1}{7} R_2$ to get

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

Step IV. Apply $R_3 \to R_3 - 5R_2$ to get

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

which is a matrix in echelon form.

Example 1.63: Reduce $A = \begin{pmatrix} 2 & 2 & 4 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix}$ to echelon form.

Solution:

Step I. Apply $R_1 \to \frac{1}{2} R_1$ to get

$$\begin{pmatrix} 1 & 1 & 2 & 2 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix}$$

Step II. Apply $R_2 \to R_2 - 2R_1, R_3 \to R_3 - 3R_1, R_4 \to R_4 - 4R_1$ to get

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

Step III. (2, 2)th place is already unity.

Step IV. Apply $R_3 \to R_3 - R_2, R_4 \to R_4 - R_2$ to get

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

Step V. Apply $R_3 \to (-1)\, R_3$ to get

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

Step VI. Apply $R_4 \to R_4 + 2R_3$ to get

$$\begin{pmatrix} 1 & 1 & 2 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

which is a matrix in echelon form.

Example 1.64: Reduce $A = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 2 & 6 & 4 \\ 0 & 3 & 9 & 3 \\ 0 & 4 & 13 & 4 \end{pmatrix}$ to echelon form.

Solution:

Step I. Since all the elements in 1st column are zero, Step I and Step II are not needed.

Step III. Apply $R_2 \to \frac{1}{2} R_2$ to get

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 1 & 3 & 2 \\ 0 & 3 & 9 & 3 \\ 0 & 4 & 13 & 4 \end{pmatrix}$$

Step IV. Apply $R_3 \to R_2 - 3R_2, R_4 \to R_4 - 4R_2$ to get

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

NOTES

Step V. Apply $R_3 \leftrightarrow R_4$ to get

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

Step VI. Since elements below (3, 3)rd place are zero, Step VI is not needed. Hence *A* is reduced to echelon form.

1.5.9 System of Linear Equations

Consider the following equations:

$$a_{11}x + a_{12}y + a_{13}z = b_{11}$$
$$a_{21}x + a_{22}y + a_{23}z = b_{12}$$
$$a_{31}x + a_{32}y + a_{33}z = b_{13}$$

These equations can be also expressed as

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} b_{11} \\ b_{12} \\ b_{13} \end{pmatrix}, \text{ i.e. } AX = B$$

where *A* is the matrix obtained by writing the coefficients of *x*, *y* and *z* in three rows respectively, and *B* is the column matrix consisting of constants in RHS of given equations.

We reduce the matrix *A* to echelon form and write the equations in the form stated earlier, and then solve the problem. A system of equations is called *consistent* if and only if there exists a common solution to all of them, otherwise it is called *inconsistent*.

Example 1.65: Solve the following system of equations:

$$x - 3y + z = -1$$
$$2x + y - 4z = -1$$
$$6x - 7y + 8z = 7$$

Solution: Let $A = \begin{pmatrix} 1 & -3 & 1 \\ 2 & 1 & -4 \\ 6 & -7 & 8 \end{pmatrix}$ and $B = \begin{pmatrix} -1 \\ -1 \\ 7 \end{pmatrix}$

and assume that there exists a matrix

$$X = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

Such that the given system of equations becomes:

$$AX = B$$

NOTES

Then, $\begin{pmatrix} 1 & -3 & 1 \\ 2 & 1 & -4 \\ 6 & -7 & 8 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ 7 \end{pmatrix}$

Apply $R_2 \to R_2 - 2R_1, R_3 \to R_3 - 6R_1$

$$\begin{pmatrix} 1 & -3 & 1 \\ 0 & 7 & -6 \\ 0 & 11 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 13 \end{pmatrix}$$

Apply $R_2 \to \frac{1}{7} R_2$

$$\begin{pmatrix} 1 & -3 & 1 \\ 0 & 1 & -\frac{6}{7} \\ 0 & 11 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ \frac{1}{7} \\ 13 \end{pmatrix}$$

Apply $R_3 \to R_3 - 11R_2$

$$\begin{pmatrix} 1 & -3 & 1 \\ 0 & 1 & -\frac{6}{7} \\ 0 & 0 & \frac{80}{7} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ \frac{1}{7} \\ \frac{80}{7} \end{pmatrix}$$

Thus, we have reduced coefficient matrix A to echelon form. Note that each elementary row operation that we applied on A was also applied on B simultaneously.

From, the last matrix equation, we have

$$x - 3y + z = -1$$

$$y - \frac{6}{7} z = \frac{1}{7}$$

$$\frac{80}{7} z = \frac{80}{7}$$

So, $z = 1, y = 1, x = 1$

Hence, the given system of equations has a solution ($x = 1, y = 1, z = 1$).

Example 1.66: Solve the following system of equations:

$$2x - 5y + 7z = 6$$

$$x - 3y + 4z = 3$$

$$3x - 8y + 11z = 11 \text{ if consistent}$$

Solution: Let $A = \begin{pmatrix} 2 & -5 & 7 \\ 1 & -3 & 4 \\ 3 & -8 & 11 \end{pmatrix}$ and $B = \begin{pmatrix} 6 \\ 3 \\ 11 \end{pmatrix}$

So, the given system of equations can be written as

$$\begin{pmatrix} 2 & -5 & 7 \\ 1 & -3 & 4 \\ 3 & -8 & 11 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \\ 11 \end{pmatrix}$$

NOTES

Apply $R_1 \leftrightarrow R_2$

$$\begin{pmatrix} 1 & -3 & 4 \\ 2 & -5 & 7 \\ 3 & -8 & 11 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \\ 11 \end{pmatrix}$$

Apply $R_2 \to R_2 - 2R_1, R_3 - 3R_1$

$$\begin{pmatrix} 1 & -3 & 4 \\ 0 & 1 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \\ 2 \end{pmatrix}$$

Apply $R_3 \to R_3 - R_2$

$$\begin{pmatrix} 1 & -3 & 4 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \\ 2 \end{pmatrix}$$

$$\Rightarrow \quad x - 3y + 4z = 3$$
$$y - z = 0$$
$$0 = 2$$

Since 0 = 2 is false, the given system of equations has no solution. So, the given system of equations is inconsistent.

Example 1.67: Solve the following system of equations:

$$x + y + z = 7$$
$$x + 2y + 3z = 16$$
$$x + 3y + 4z = 22$$

Solution: The given system of equations in matrix form:

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 7 \\ 16 \\ 22 \end{pmatrix}$$

Apply $R_2 \to R_2 - R_1, R_3 \to R_3 - R_1$

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 2 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 7 \\ 9 \\ 15 \end{pmatrix}$$

Apply $R_3 \to R_3 - 2R_1$

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 7 \\ 9 \\ -3 \end{pmatrix}$$

$$\Rightarrow \quad x + y + z = 7$$
$$y + 2z = 9$$
$$z = 3$$

So, $z = 3, y = 3, x = 1$

The given system of equations has a solution (1, 3, 3).

NOTES

Example 1.68: Solve the following system of equations:

$$x + y + z = 2$$
$$x + 2y + 3z = 5$$
$$x + 3y + 6z = 11$$
$$x + 4y + 10z = 21$$

Solution: We have

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 6 \\ 1 & 4 & 10 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ 11 \\ 21 \end{pmatrix}$$

Apply $R_2 \to R_2 - R_1, R_3 \to R_3 - R_1, R_4 \to R_4 - R_1$

Then
$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 2 & 5 \\ 0 & 3 & 9 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 9 \\ 19 \end{pmatrix}$$

Apply $R_3 \to R_3 - 2R_2, R_4 \to R_4 - 3R_2$

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 3 \\ 10 \end{pmatrix}$$

Apply $R_4 \to R_4 - 3R_3$

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 3 \\ 1 \end{pmatrix}$$

$\Rightarrow$ $x + y + z = 2, y + 2z = 3, z = 3, 0 = 1$

This is absurd. So, the given system is inconsistent.

Example 1.69: Solve

$$x - 3y - 3z = -10$$
$$3x + y - 4z = 0$$
$$2x + 5y + 6z = 13$$

Solution: We have

$$\begin{pmatrix} 1 & -3 & -8 \\ 3 & 1 & -4 \\ 2 & 5 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -10 \\ 0 \\ 13 \end{pmatrix}$$

Apply $R_2 \to R_2 - 3R_1, R_3 \to R_3 - 2R_1$

$$\begin{pmatrix} 1 & -3 & -8 \\ 0 & 10 & 20 \\ 0 & 11 & 22 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -10 \\ 30 \\ 33 \end{pmatrix}$$

NOTES

Apply $R_2 \to \frac{1}{10} R_2$

$$\begin{pmatrix} 1 & -3 & -8 \\ 0 & 1 & 2 \\ 0 & 11 & 22 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -10 \\ 3 \\ 33 \end{pmatrix}$$

Apply $R_3 \to R_3 - 11R_2$

$$\begin{pmatrix} 1 & -3 & -8 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -10 \\ 3 \\ 0 \end{pmatrix}$$

$\Rightarrow$ $x - 3y - 8z = -10$

$y + 2z = 3$

Let $z = k \Rightarrow y = 3 - 2k$

and $x = 9 - 6k + 8k - 10 = 2k - 1$

So, the given system has infinite number of solutions of the form $x = 2k - 1$, $y = 3 - 2k$, $z = k$ where k is any number.

Example 1.70: Solve

$$x + 2y + 3z + 4w = 0$$
$$8x + 5y + z + 4w = 0$$
$$5x + 6y + 8z + w = 0$$
$$8x + 3y + 7z + 2w = 0$$

Solution: We have

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 8 & 5 & 1 & 4 \\ 5 & 6 & 8 & 1 \\ 8 & 3 & 7 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Apply $R_2 \to R_2 - 8R_1$, $R_3 \to R_3 - 5R_1$, $R_4 \to R_4 - 8R_1$

$\Rightarrow$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -11 & -23 & -28 \\ 0 & -4 & -7 & -19 \\ 0 & -13 & -17 & -30 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Apply $R_2 \to -\frac{1}{11} R_2$

$\Rightarrow$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & \frac{23}{11} & \frac{28}{11} \\ 0 & -4 & -7 & -19 \\ 0 & -13 & -17 & -30 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

NOTES

Apply $R_3 \to R_3 + 4R_2,\ R_4 \to R_4 + 13R_2$

$$\Rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & \frac{23}{11} & \frac{28}{11} \\ 0 & 0 & \frac{15}{11} & -\frac{97}{11} \\ 0 & 0 & \frac{112}{11} & \frac{34}{11} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Apply $R_3 \to \frac{11}{15} R_3$

$$\Rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & \frac{23}{11} & \frac{28}{11} \\ 0 & 0 & 1 & -\frac{97}{15} \\ 0 & 0 & \frac{112}{11} & \frac{34}{11} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Apply $R_4 = -\frac{112}{11} R_3$

$$\Rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & \frac{23}{11} & \frac{28}{11} \\ 0 & 0 & 1 & \frac{97}{15} \\ 0 & 0 & 0 & \frac{11374}{165} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\Rightarrow \quad x + 2y + 3z + 4w = 0$$

$$y + \frac{23}{11} z + \frac{28}{11} w = 0$$

$$z - \frac{97}{15} w = 0$$

$$w = 0$$

$$\Rightarrow \quad x = 0,\ y = 0,\ z = 0,\ w = 0.$$

Thus, the system has only one solution, namely $x = y = z = w = 0$.

1.5.10 Inverse of a Matrix

Consider the following matrices:

$$A = \begin{pmatrix} 2 & 0 & -1 \\ 5 & 1 & 0 \\ 0 & 1 & 3 \end{pmatrix} \text{ and } B = \begin{pmatrix} 3 & -1 & 1 \\ -15 & 6 & -5 \\ 5 & -2 & 2 \end{pmatrix}$$

It can be easily seen that

$$AB = BA = I \text{ (unit matrix)}$$

In this case, we say B is an inverse of A. In fact, we have the following definition.

If A is a square matrix of order n, then a square matrix B of the same order n is said to be an inverse of A if $AB = BA = I$ (unit matrix).

NOTES

Notes: (*i*) Inverse of a matrix is defined only for square matrices.

(*ii*) If B is an inverse of A, then A is also an inverse of B. (Follows clearly by definition.)

(*iii*) If a matrix A has an inverse, then A is said to be invertible.

(*iv*) Inverse of a matrix is unique.

For example, let B and C be two inverses of A.

Then, $AB = BA = I$ and $AC = CA = I$

So, $B = BI = B(AC) = (BA)C = IC = C$

Notation: Inverse of A is denoted by A^{-1}.

(*v*) Every square matrix is not invertible.

For example, let $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$

If A is invertible, let $B = \begin{pmatrix} x & x' \\ y & y' \end{pmatrix}$ be inverse of A.

Then $AB = I$ implies $\begin{pmatrix} x+y & x'+y' \\ x+y & x'+y' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$\Rightarrow x + y = 1,\ x' + y' = 0,\ x + y = 0,\ x' + y' = 1$ which is absurd.

This proves our assertion.

Let us discuss the method to determine the inverse of a matrix. Consider the identity $A = IA$.

We reduce the matrix A on left hand side to the unit matrix I by elementary row operations only and apply all those operations in the same order to the prefactor I on the right hand side of the above identity. In this way, unit matrix I is reduced to same matrix B such that $I = BA$. Matrix B is then the inverse of A.

We illustrate the preceding method by the following examples.

Example 1.71: Find the inverse of the following matrix:

$$\begin{pmatrix} 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 3 & 4 \end{pmatrix}$$

Solution: Consider the identity

$$\begin{pmatrix} 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 3 & 4 \end{pmatrix}$$

Applying $R_2 \to R_2 - R_1$, then $R_3 \to R_3 - R_1$, we have

$$\begin{pmatrix} 1 & 3 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 3 & 4 \end{pmatrix}$$

Applying $R_1 \to R_1 - 3R_2 - 3R_3$, we have

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 7 & -3 & -3 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 3 & 4 \end{pmatrix}$$

NOTES

So, the desired inverse is:

$$\begin{pmatrix} 7 & -3 & -3 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

Example 1.72: Find the inverse of the following matrix:

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

Solution: Consider the following identity:

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

Applying $R_2 \to R_2 + 3R_1$, $R_3 \to R_3 - 2R_1$, we have

$$\begin{pmatrix} 1 & 3 & -2 \\ 0 & 9 & -11 \\ 0 & -1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 3 & -2 \\ -3 & 0 & -5 \\ 2 & 5 & 0 \end{pmatrix}$$

Applying $R_3 \to 9R_3$ and then $R_3 \to R_3 + R_2$, we have

$$\begin{pmatrix} 1 & 3 & -2 \\ 0 & 9 & -11 \\ 0 & 0 & 25 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -15 & 1 & 9 \end{pmatrix} \begin{pmatrix} 1 & 3 & -2 \\ -3 & 0 & -5 \\ 2 & 5 & 0 \end{pmatrix}$$

Applying $R_3 \to \frac{1}{25} R_3$, we have

$$\begin{pmatrix} 1 & 3 & -2 \\ 0 & 9 & -11 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -\frac{3}{5} & \frac{1}{25} & \frac{9}{25} \end{pmatrix} \begin{pmatrix} 1 & 3 & -2 \\ -3 & 0 & -5 \\ 2 & 5 & 0 \end{pmatrix}$$

Applying $R_2 \to R_2 + 11R_3$, $R_1 \to R_1 + 2R_3$, we have

$$\begin{pmatrix} 1 & 3 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{5} & \frac{2}{25} & \frac{18}{25} \\ -\frac{18}{5} & \frac{36}{25} & \frac{99}{25} \\ -\frac{3}{5} & \frac{1}{25} & \frac{9}{25} \end{pmatrix} \begin{pmatrix} 1 & 3 & -2 \\ -3 & 0 & -5 \\ 2 & 5 & 0 \end{pmatrix}$$

Applying $R_2 \to \frac{1}{9} R_2$, we have

$$\begin{pmatrix} 1 & 3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{5} & \frac{2}{25} & \frac{18}{25} \\ -\frac{2}{5} & \frac{4}{25} & \frac{11}{25} \\ -\frac{3}{5} & \frac{1}{25} & \frac{9}{25} \end{pmatrix} \begin{pmatrix} 1 & 3 & -2 \\ -3 & 0 & -5 \\ 2 & 5 & 0 \end{pmatrix}$$

Applying $R_1 \to R_1 - 3R_2$, we have

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{2}{5} & \frac{3}{5} \\ -\frac{2}{5} & \frac{4}{25} & \frac{11}{25} \\ -\frac{3}{4} & \frac{1}{25} & \frac{9}{25} \end{pmatrix} \begin{pmatrix} 1 & 3 & -2 \\ -3 & 0 & -5 \\ 2 & 5 & 0 \end{pmatrix}$$

So, the desired inverse is:

$$\begin{pmatrix} 1 & -\frac{2}{5} & -\frac{3}{5} \\ -\frac{2}{5} & \frac{4}{25} & \frac{11}{25} \\ -\frac{3}{5} & \frac{1}{25} & \frac{9}{25} \end{pmatrix}$$

Example 1.73: Find the inverse of the following matrix:

$$\begin{pmatrix} 1 & 2 & -1 \\ -4 & -7 & 4 \\ -4 & -9 & 5 \end{pmatrix}$$

Solution: Consider the following identity:

$$\begin{pmatrix} 1 & 2 & -1 \\ -4 & -7 & 4 \\ -4 & -9 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & -1 \\ -4 & -7 & 4 \\ -4 & -9 & 5 \end{pmatrix}$$

Applying $R_2 \to R_2 + 4R_1$, $R_3 \to R_3 + 4R_1$, we have

$$\begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 4 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & -1 \\ -4 & -7 & 4 \\ -4 & -9 & 5 \end{pmatrix}$$

Applying $R_1 \to R_1 + R_3$ then $R_3 \to R_3 + R_2$:

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 0 & 1 \\ 4 & 1 & 0 \\ 8 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & -1 \\ -4 & -7 & 4 \\ -4 & -9 & 5 \end{pmatrix}$$

Applying $R_1 \to R_1 - R_2$, we have

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 4 & 1 & 0 \\ 8 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & -1 \\ -4 & -7 & 4 \\ -4 & -9 & 5 \end{pmatrix}$$

So, the desired inverse is:

$$\begin{pmatrix} 1 & -1 & 1 \\ 4 & 1 & 0 \\ 8 & 1 & 1 \end{pmatrix}$$

1.5.11 Rank of a Matrix

Suppose we have a 3×4 matrix:

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}$$

If we delete any one column from it, we get corresponding 3×3 submatrix. The determinant of any one of these is called a minor of the matrix *A*. Thus,

$\begin{vmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \\ 9 & 10 & 11 \end{vmatrix}$, $\begin{vmatrix} 1 & 2 & 4 \\ 5 & 6 & 8 \\ 9 & 10 & 12 \end{vmatrix}$, $\begin{vmatrix} 1 & 3 & 4 \\ 5 & 7 & 8 \\ 9 & 11 & 12 \end{vmatrix}$ and $\begin{vmatrix} 2 & 3 & 4 \\ 6 & 7 & 8 \\ 10 & 11 & 12 \end{vmatrix}$ are called minors of *A*.

These are 3×3 determinants (sometimes called 3-rowed minors).

Similarly, if we delete any one row and two columns of *A*, we get the corresponding 2-rowed minors.

Definition: Let *A* be an $m \times n$ matrix. We say rank of *A* is *r* if (*i*) at least one minor of order *r* is non-zero and (*ii*) every minor of order $(r + 1)$ is zero.

Example 1.74: Find the rank of the following matrix:

$$\begin{pmatrix} 1 & 2 & 4 \\ -1 & -2 & -4 \\ 2 & 4 & 8 \\ -3 & -6 & -9 \end{pmatrix}$$

Solution: Since it is a 3×4 matrix, it cannot have a minor with order larger than 3×3.

Now 3×3 minors of this matrix are:

$\begin{vmatrix} 1 & 2 & 4 \\ -1 & -2 & -4 \\ 2 & 4 & 8 \end{vmatrix}$, $\begin{vmatrix} 1 & 2 & 4 \\ -1 & -2 & -4 \\ -3 & -6 & -9 \end{vmatrix}$, $\begin{vmatrix} 1 & 2 & 4 \\ 2 & 4 & 8 \\ -3 & -6 & -9 \end{vmatrix}$, $\begin{vmatrix} -1 & -2 & -4 \\ 2 & 4 & 8 \\ -3 & -6 & -9 \end{vmatrix}$

You can check that all these are zero. So rank of *A* is less than 3.

Again since a 2×2 minor $\begin{vmatrix} 4 & 8 \\ -6 & -9 \end{vmatrix} \neq 0$.

We find rank is $\geq$, 2, i.e. rank of *A* is 2.

Example 1.75: Find the rank of the following matrix:

$$A = \begin{vmatrix} 0 & i & -i \\ -i & 0 & i \\ i & -i & 0 \end{vmatrix}$$

NOTES

Solution: We have

$$|A| = \begin{vmatrix} 0 & i & -i \\ -i & 0 & i \\ i & -i & 0 \end{vmatrix} = 0$$

Thus, the rank $A \leq 2$.

Again as $\begin{vmatrix} 0 & -i \\ -i & 0 \end{vmatrix} = -1 \neq 0$

rank $A \geq 2$ and hence rank $A = 2$.

Notes: (*i*) It is easy ot see that if the given matrix A is $m \times n$ matrix, then rank $A \leq \min(m, n)$.

(*ii*) If in A, every $r \times r$ determinant is zero, then the rank is less than or equal to r–1.

(*iii*) If $\exists$ a non-zero $r \times r$ determinant, then the rank is greater than or equal to r.

(*iv*) Rank of null matrix is taken as zero.

(*v*) If every r-rowed minor is zero, then every higher order minor would automatically be zero.

You can prove that the rank of a matrix remains unchanged by elementary operations. In view of this result, the process of finding a rank can be simplified. We first reduce the given matrix to a triangular form by elementary row operations and then find the rank of the new matrix which is the rank of the original matrix.

Example 1.76: Find the rank of the following matrix:

$$A = \begin{pmatrix} 1 & -3 & 2 \\ -3 & 9 & -6 \\ 2 & -6 & 4 \end{pmatrix}$$

Solution: We have

$$\text{A} = \begin{pmatrix} 1 & -3 & 2 \\ -3 & 9 & -6 \\ 2 & -6 & 4 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 \\ -3 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix} \quad \text{using} \quad C_2 \to C_2 + 3C_1, \; C_3 \to C_3 - 2C_1$$

$$\sim \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{using} \quad R_2 \to R_2 + 3R_1, \; R_3 \to R_3 - 2R_1$$

So, rank of A is 1.

Example 1.77: Find the rank of the following matrix:

$$A = \begin{pmatrix} 2 & 3 & -1 & -1 \\ 1 & -1 & -2 & -4 \\ 6 & 1 & 3 & -2 \\ 6 & 3 & 0 & -7 \end{pmatrix}$$

Solution: We have

$$A \sim \begin{pmatrix} 2 & 3 & -1 & -1 \\ 1 & -1 & -2 & -4 \\ 6 & 1 & 3 & -2 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & -2 & -4 \\ 2 & 3 & -1 & -1 \\ 6 & 1 & 3 & -2 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & -2 & -4 \\ 0 & 5 & 3 & 7 \\ 0 & 7 & 15 & 22 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$R_4 \to R_4 - R_3 - R_2 - R_1$, $R_1 \leftrightarrow R_2$, $R_2 \to R_2 - 2R_1$ and $R_3 \to R_3 - 6R_1$

$$\sim \begin{pmatrix} 1 & -2 & -2 & -4 \\ 0 & 35 & 21 & 49 \\ 0 & 35 & 75 & 110 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & -2 & -4 \\ 0 & 35 & 21 & 49 \\ 0 & 0 & 54 & 61 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Here $\begin{vmatrix} 1 & -1 & -2 \\ 0 & 35 & 21 \\ 0 & 0 & 54 \end{vmatrix} \neq 0$

Hence, the rank of A is 3.

Example 1.78: Find the rank of the following matrix:

$$A = \begin{pmatrix} 5 & 3 & 14 & 4 \\ 0 & 1 & 2 & 1 \\ 1 & -1 & 2 & 0 \end{pmatrix}$$

Solution: Apply $R_1 \leftrightarrow R_3$, then

$$\text{A} \sim \begin{pmatrix} 1 & -1 & 2 & 0 \\ 0 & 1 & 2 & 1 \\ 5 & 3 & 14 & 4 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & 2 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 8 & 4 & 4 \end{pmatrix} \quad \text{by } R_3 \to R_3 - 5R_1$$

$$\sim \begin{pmatrix} 1 & -1 & 2 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & -12 & -4 \end{pmatrix} \quad \text{using } R_3 \to R_3 - 8R_2$$

Since this reduced matrix has non-zero 3-rowed minor

$$\begin{vmatrix} 1 & -1 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & -12 \end{vmatrix}$$

its rank is 3. Also, there is no 4-rowed minor. Hence, the rank of given matrix is also 3.

1.5.12 Business Applications of Matrices

Application of matrices to business can be understood with the help of following examples

NOTES

Example 1.79: In an examination of mathematics, 20 students from college *A*, 30 students from college *B* and 40 students from college *C* appeared. Only 15 students from each college could get through the examination. Out of them 10 students from college *A*, 5 students from college *B* and 10 students from college *C* secured full marks. Write down the data in matrix form.

Solution: Consider the following matrix:

$$\begin{pmatrix} 20 & 30 & 40 \\ 15 & 15 & 15 \\ 10 & 5 & 10 \end{pmatrix}$$

The first row represents the number of students in college *A*, college *B*, college *C* respectively. The second row represents the number of students who got through the examination in three colleges respectively. The third row represents the number of students who got full marks in the three colleges respectively.

Example 1.80: A publishing house has two branches. In each branch, there are three offices. In each office, there are 3 peons, 4 clerks and 5 typists. In one office of a branch, 6 salesmen are also working. In each office of the other branch 2 head clerks are also working. Using matrix notation, find: (i) the total number of posts of each kind in all the offices taken together in each branch and (ii) the total number of posts of each kind in all the offices taken together from both the branches.

Solution: (*i*) Consider the following row matrices:

$$A_1 = (3\ \ 4\ \ 5\ \ 6\ \ 0), \quad A_2 = (3\ \ 4\ \ 5\ \ 0\ \ 0), \quad A_3 = (3\ \ 4\ \ 5\ \ 0\ \ 0)$$

These matrices represent the three offices of one branch (say *A*) where elements appearing in the row represent the number of peons, clerks, typists, salesmen and head clerks taken in that order working in the three offices.

Then, $A_1 + A_2 + A_3 = (3+3+3\ \ 4+4+4\ \ 5+5+5\ \ 6+0+0\ \ 0+0+0)$

$= (9\ \ 12\ \ 15\ \ 6\ \ 0)$

Thus, the total number of posts of each kind in all the offices of branch *A* are the elements of matrix $A_1 + A_2 + A_3 = (9\ \ 12\ \ 15\ \ 6\ \ 0)$

Now, consider the following row matrices:

$$B_1 = (3\ \ 4\ \ 5\ \ 0\ \ 2), \quad B_2 = (3\ \ 4\ \ 5\ \ 0\ \ 2), \quad B_3 = (3\ \ 4\ \ 5\ \ 0\ \ 2)$$

Here B_1, B_2, B_3 represent three offices of other branch (say *B*) where the elements in the row represents number of peons, clerks, typists, salesmen and head clerks respectively.

Thus, the total number of posts of each kind in all the offices of branch *B* are the elements of the matrix $B_1 + B_2 + B_3 = (9\ \ 12\ \ 15\ \ 0\ \ 6)$

(*ii*) The total number of posts of each kind in all the offices taken together from both branches are the elements of the following matrix:

$$(A_1 + A_2 + A_3) + (B_1 + B_2 + B_3) = (18 \;\; 24 \;\; 30 \;\; 6 \;\; 6)$$

NOTES

Example 1.81: Let $A = \begin{pmatrix} 10 & 20 \\ 30 & 40 \end{pmatrix}$

where the first row represents the number of table fans and the second row represents the number of ceiling fans which two manufacturing units *A* and *B* make in one day. The first and second columns represent the manufacturing units *A* and *B*. Compute 5*A* and state what it represents.

Solution: $5A = \begin{pmatrix} 50 & 100 \\ 150 & 200 \end{pmatrix}$

It represents the number of table fans and ceiling fans that the manufacturing units *A* and *B* produce in five days.

Example 1.82: Let $A = \begin{pmatrix} 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix}$

where rows represent the number of items of types I, II and III, respectively. The four columns represents the four shops A_1, A_2, A_3 and A_4, respectively.

Let $B = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \end{pmatrix}$ and $C = \begin{pmatrix} 1 & 2 & 2 & 3 \\ 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 4 \end{pmatrix}$

where elements in matrix *B* represent the number of items of different types delivered at the beginning of a week and those in matrix *C* represent the sales during that week. Find:

(*i*) The number of items immediately after delivery of items.

(*ii*) The number of items at the end of the week.

(*iii*) The number of items needed to bring stocks of all items in all shops to 6.

Solution: (*i*) $A + B = \begin{pmatrix} 3 & 5 & 7 & 9 \\ 5 & 5 & 7 & 9 \\ 7 & 7 & 7 & 9 \end{pmatrix}$

It represents the number of items immediately after delivery of items.

(*ii*) $(A + B) - C = \begin{pmatrix} 2 & 3 & 5 & 6 \\ 4 & 3 & 4 & 5 \\ 5 & 4 & 3 & 5 \end{pmatrix}$

It represents the number of items at the end of the week.

(*iii*) You want that all elements in $(A + B) - C$ should be 6.

Let $D = \begin{pmatrix} 4 & 3 & 1 & 0 \\ 2 & 3 & 2 & 1 \\ 1 & 2 & 3 & 1 \end{pmatrix}$

Then $(A + B) - C + D$ is a matrix in which all elements are 6. So, *D* represents the number of items needed to bring stocks of all items of all shops to 6.

NOTES

Example 1.83: The following matrix represents the results of the examination of B. Com. class:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}$$

The rows represent the three sections of the class. The first three columns represent the number of students securing 1st, 2nd and 3rd divisions, respectively, in that order and the fourth column represents the number of students who failed in the examination in that order.

Answer the following quesions:

(*i*) How many students passed in three sections respectively?

(*ii*) How many students failed in three sections respectively?

(*iii*) Write down the matrix showing the number of successful students.

(d) Write down the column matrix showing only failed students.

(e) Write down the column matrix showing students who have secured 1st division from three sections.

Solution: (*i*) The number of students who passed in three sections respectively are 1 + 2 + 3 = 6, 5 + 6 + 7 = 18, 9 + 10 + 11 = 30.

(*ii*) The numbers of students who failed from three sections are 4, 8 and 12, respectively.

(*iii*) $\begin{pmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \\ 9 & 10 & 11 \end{pmatrix}$

(*iv*) $\begin{pmatrix} 4 \\ 8 \\ 12 \end{pmatrix}$ represents the column matrix where only failed students are shown.

(*v*) $\begin{pmatrix} 1 \\ 5 \\ 9 \end{pmatrix}$ represents the column matrix of students securing 1st division.

Example 1.84: In South Delhi there are 20 colleges and 50 schools. Each of them has 1 peon, 5 clerks, 1 cashier. Each college, in addition, has 1 accountant and 1 head clerk. The monthly salary of each of them is as follows. Peon Rs 150, clerk Rs 250, cashier Rs 300, accountant Rs 350 and head clerk Rs 400. Using matrix notation, find:

(*i*) The total number of posts of each kind in schools and colleges taken together.

(*ii*) The total monthly salary bill of each school and college separately.

(*iii*) The total monthly salary bill of all the schools and colleges taken together.

Solution: (*i*) Consider the row matrix of order 1 × 2

$$A = (20 \quad 50)$$

NOTES

This represents the number of colleges and schools in that order.

Let $B = \begin{pmatrix} 1 & 5 & 1 & 1 & 1 \\ 1 & 5 & 1 & 0 & 0 \end{pmatrix}$

where columns represent the numbers of peons, clerks, cashier, accountants, headclerks while rows represent the numbers of colleges and schools in that order.

Then $AB = (70 \quad 350 \quad 70 \quad 20 \quad 20)$

where the first element represents the total number of peons, the second represents the total number of clerks, the third represents the total number of cashier, the fourth represents the total number of accountants and the fifth element represents the total number of head clerks.

(*ii*) Let $C = \begin{pmatrix} 150 \\ 250 \\ 300 \\ 350 \\ 400 \end{pmatrix}$

where column matrix represents the monthly salary of peon, clerk, cashier, accountant and head-clerk in that order.

Then $BC = \begin{pmatrix} 150 + 1250 + 300 + 350 + 400 \\ 150 + 1250 + 300 + 0 + 0 \end{pmatrix}$

$= \begin{pmatrix} 2450 \\ 1700 \end{pmatrix}$

Thus, the total monthly salary bill of each college is Rs 2450 and of each school is Rs 1700.

(*iii*) Consider $A(BC) = (2450 \times 20 + 50 \times 1700)$

$= (49{,}000 + 85{,}000)$

$= (134{,}000)$

which gives the total monthly salary bill of schools and colleges taken together.

Example 1.85: A manufacturing unit produces three types of television sets *A, B, C*. The following matrix shows the sale of television sets in two different cities:

$$\begin{pmatrix} A & B & C \\ 400 & 300 & 200 \\ 300 & 200 & 100 \end{pmatrix}$$

If the cost price of each set *A, B, C* is Rs 1000, Rs 2000 and Rs 3000 respectivley and the selling price is Rs 1500, Rs 3000 and Rs 4000 respectively, find the total profit using matrix algebra only.

Solution: Consider the product

$$(1000 \quad 2000 \quad 3000) \begin{pmatrix} 400 & 300 \\ 300 & 200 \\ 200 & 100 \end{pmatrix}$$

NOTES

$= (400 \times 1000 + 300 \times 2000 + 200 \times 3000 \quad 300 \times 1000 + 200 \times 2000 + 100 \times 3000)$

$= (400{,}000 + 600{,}000 + 600{,}000 \quad 300{,}000 + 400{,}000 + 300{,}000)$

$= (1{,}60{,}0000 + 1{,}000{,}000)$

Thus, total money spent in producing the television sets is Rs (1600000 + 1,000,000) is 2,600,000. Consider the product.

$$(1500 \quad 3000 \quad 4000)\begin{pmatrix} 400 & 300 \\ 300 & 200 \\ 200 & 100 \end{pmatrix}$$

$= (600{,}000 + 900{,}000 + 800{,}000 \quad 450{,}000 + 600{,}000 + 400{,}000)$

$= (2{,}300{,}000 \quad 14{,}500{,}000)$

Hence, the total money obtained after selling the sets is Rs 23,00,000 + 14,50,000 Rs 37,50,000 and the total profit is Rs 11,50,000.

Example 1.86: The cost of manufacturing three types of motor cars is given below:

Car	Labour hrs.	Material used	Subcontracted work
A	40	100 units	50 units
B	80	150 units	80 units
C	100	250 units	100 units

Labour cost is £ 2 per hour, per unit material cost is 10 $ and one unit of subcontracted work costs £1. Find the total cost of manufacturing 3000, 2000 and 1000 vehicles of type *A*, *B*, *C*, respectively.

Solution: Consider the following matrices:

$$M = \begin{pmatrix} 40 & 100 & 50 \\ 80 & 150 & 80 \\ 100 & 250 & 100 \end{pmatrix} \text{ and } N = \begin{pmatrix} 2 \\ 1/2 \\ 1 \end{pmatrix}$$

M represents labour hours, material used and subcontracted work for three types of cars, *A*, *B*, *C*, respectively.

Column matrix *N* represents the labour cost per unit material cost and cost of subcontracted work.

$$\text{Then } MN = \begin{pmatrix} 180 \\ 315 \\ 425 \end{pmatrix}$$

The column matrix represents cost of each car *A*, *B*, *C* in that order.

Let $P = (3000 \quad 2000 \quad 1000)$

This row matrix *P* represents the number of cars *A*, *B*, *C* to be manufactured in that order.

Consider $P(MN) = (15{,}95{,}000)$

Thus total cost of manufacturing three cars *A*, *B*, *C* is £15,95,000.

Example 1.87: The following matrix shows the sale of cold drinks in a shop during one week from Monday to Sunday.

$$A = \begin{matrix} C & F & L \\ 20 & 25 & 30 \\ 25 & 30 & 40 \\ 30 & 25 & 20 \\ 40 & 30 & 50 \\ 45 & 40 & 20 \\ 50 & 20 & 30 \\ 60 & 40 & 60 \end{matrix}$$

where C = Coca-Cola, F = Fanta and L = Limca. The cost of each bottle of C, F, L is Re 1, Rs 2 and Rs 3, respectively. Using matrix algebra only:

(*i*) Find the total sales of C, F and L separately during one week,

(*ii*) Find the total earning during one week.

(*iii*) Find total sale of C, F and L taken together each day, from Monday to Sundary.

(*iv*) Find the total sale of C, F and L taken together during one week.

Solution: (*i*) Consider the product

$$(1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1)\ A = (270 \quad 210 \quad 250)$$

This represents the total sales of C, F and L separately during one week.

(*ii*) Consider the product

$$(270 \quad 210 \quad 250) \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \quad (270 + 420 + 750) = (1440)$$

Thus, the total earning during one week is Rs 1440.

(*iii*) Consider the product

$$A \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = (75\ \ 95\ \ 75\ \ 120\ \ 105\ \ 100\ \ 160)$$

This gives the total sale of C, F and L taken together each day, from Monday to Sunday.

(*iv*) Consider the product

$$(1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1)\, A \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = (270\ \ 210\ \ 250) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$= (730)$$

Thus, the total sales of C, F and L taken together during one week are of 730 bottles.

CHECK YOUR PROGRESS

9. Define matrix.
10. Name the various types of matrices.
11. What are the types of an elementary operation?

1.6 SUMMARY

In this unit, you have learned that:

- In mathematics, a function expresses the inherent idea that one quantity (input) completely determines another quantity (output).
- There are various types of functions, such as one-to-one function, onto function, bijective function, constant function, into function and identity function.
- There are two types of progressions in mathematics: arithmetic progression and geometric progression.
- The **arithmetic progression** is a sequence of numbers in which the difference of any two successive members of the sequence is a constant.
- The geometric progression refers to a sequence of numbers in which each number is obtained from the previous one by multiplying it by a constant.
- The geometric mean refers to a type of mean or average which indicates the central tendency or typical value of a set of numbers.
- A matrix (plural matrices**)** is a rectangular array of numbers.
- There are various types of matrices, such as row matrix, column matrix, square matrix, null or zero matrix, diagonal matrix, scalar matrix, identity matrix and triangular matrix.

1.7 KEY TERMS

- **Arithmetic progression:** It is a sequence of numbers in which the difference of any two successive members of the sequence is a constant.
- **Geometric progression:** It is a sequence of numbers in which each number is obtained from the previous one by multiplying by a constant.
- **Matrix:** It is rectangular array of numbers.
- **Row matrix:** It is a matrix which has exactly one row.
- **Column matrix:** It is a matrix which has exactly one column.

NOTES

- **Square matrix:** It is a matrix in which the number of rows is equal to the number of columns.
- **Null or zero matrix**: It is a matrix each of whose elements is zero.
- **Scalar matrix***:* It is a diagonal matrix whose diagonal elements are equal.
- **Identity matrix**: It is a diagonal matrix whose diagonal elements are all equal to 1 (unity).

1.8 ANSWERS TO 'CHECK YOUR PROGRESS'

1. In mathematics, a function expresses the inherent idea that one quantity (input) completely determines another quantity (output). A function assigns a unique value to each input of a specified type.
2. The various types of functions are: one-to-one function, onto function, bijective function, constant function, into function and identity function.
3. The ceiling function assigns to the real number x the smallest integer that is greater than or equal to x. The value of this function is denoted by $\lceil x \rceil$.
4. The most common hashing function is the function $h(p) = p \bmod q$, where q is the number of available memory locations.
5. The arithmetic progression is a sequence of numbers in which the difference of any two successive members of the sequence is a constant.
6. The arithmetic mean of a list of numbers refers to the total of the entire list divided by the number of items in the list.
7. The geometric progression refers to a sequence of numbers in which each number is obtained from the previous one by multiplying it by a constant.
8. The geometric mean refers to the type of mean or average which indicates the central tendency or typical value of a set of numbers.
9. A matrix is a rectangular array of numbers.
10. The various types of matrices are: row matrix, column matrix, square matrix, null or zero matrix, diagonal matrix, scalar matrix, identity matrix and triangular matrix.
11. An elementary operation is either elementary row operation or elementary column operation and is of the following three types:

    Type I. The interchange of any two rows (or column).

    Type II. The multiplication of any row (or column) by a non-zero number.

    Type III. The addition of multiple of one row (or column) to another row (or column).

NOTES

1.9 QUESTIONS AND EXERCISES

Short-Answer Questions

1. Find whether each of the following functions from {1, 2, 3, 4} to itself is 1–1.
   (a) $f(1)=2; f(2)=1; f(3)=3; f(4)=4$
   (b) $f(1)=2; f(2)=2; f\{3)=4; f(4)=3$
   (c) $f(1)=4; f(2)=2; f(3)=3; f(4)=4$
2. Find whether each of the following functions is a bijection from R to R.
   (a) $f(x) = -3+4$ (b) $f(x) = -3x^2+7$
   (c) $f(x) = (x+1)/(x+2)$ (d) $f(x) = x^5 + 1$
   (e) $f(x) = 2x + 1$ (f) $f(x) = x^2 + 1(g)f(x) = x^3$
3. Let $A = \{-1, 0, 2, 4, 9\}$. Find $f(s)$ if
   (a) $f(x) = 1$ (b) $f(x)=3x+1$
   (c) $f(x) = \left[\frac{x}{2}\right]$
4. If f and $f \circ g$ are 1–1, does it follow that g is 1–1? Justify your answer.
5. Find (i) $f \circ g$, (ii) $g \circ f$, (iii) $f \circ f$ and (iv) $g \circ g$ where $f(x) = x^2 + 3x + 2$ and $g(x) = 2x-3$ are functions from R to R.
6. Find the inverse of the function $f(x) = x^3 + 1$.
7. The third and 13th term of an AP are respectively equal to –40 and 0. Find the AP and its 20th term.
8. Is 203 any term of AP, 13, 28, 43, 58, 73,...?
9. Find the sum of the following series:
   (*i*) 49, 44, 39,... up to 17 terms.
   (*ii*) 3.75, 3.5, 3.25,... up to 16 terms.
   (*iii*) $2a - b, 4a - 3b, 6a - 3b$, ... up to n terms.
   (*iv*) $5\frac{1}{2}, 6\frac{3}{4}$, 8, ... up to 28 terms.
   (*v*) $\frac{3}{\sqrt{5}}, \frac{4}{\sqrt{5}}, \sqrt{5}$, ... up to 25 terms.
10. Insert nine arithmetic means between $\frac{1}{4}$ and $\frac{-39}{4}$.
11. The third term of an AP is 18 and seventh term is 30. Find the sum of 17 terms.
12. The sum of four integers in AP is 24 and their product is 945. Find them.
    (*Hint.* Select the numbers to be $a - 3d, a - d, a + d$ and $a + 3d$.)

13. Find the sum of following series:

(*i*) $\frac{1}{2}+\frac{1}{3}+\frac{2}{9}\cdots$ up to 7 terms.

(*ii*) $16.2 + 5.4 + 1.8 + \cdots$ up to 9 terms.

(*iii*) $-\frac{1}{3}+\frac{1}{2}-\frac{3}{4}+\cdots$ up to 10 terms.

(*iv*) $3 + \sqrt{3} + 1 + \cdots$ up to 4 terms and infinity.

(*v*) $1.665 + (-1.11) + .74 + \cdots$ up to infinity.

14. (*i*) Insert five geometric means between $\frac{32}{9}$ and $\frac{81}{2}$.

(*ii*) Insert three geometric means between $\frac{9}{4}$ and $\frac{4}{9}$.

15. (*a*) Sum the following series:

(*i*) $1 + 4x + 7x^2 + 10x^3 + \cdots$ up to infinity, $x < 1$.

(*ii*) $3 + 33 + 333 + \cdots$ up to 8 terms.

(*b*) Evaluate

(*i*) .132 (*ii*) .178.

16. If *p*th, *q*th, *r*th term of a GP are, respectively, equal to *a*, *b*, *c*, then prove that

$$a^{q-r}\, b^{r-p}\, c^{p-q} = 1.$$

17. If *a*, *b*, *c*, *d* are in GP, show that

(*i*) $(a-b)^2$, $(b-c)^2$ and $(c-d)^2$ are in GP;

(*ii*) $a^2 + b^2$, $ab + bc$ and $b^2 + c^2$ are in GP;

(*iii*) $(b-c)^2 + (c-a)^2 + (d-b)^2 = (a-d)^2$.

18. If $A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 3 \\ 4 & 2 & 1 \end{pmatrix}$ find $a_{11}, a_{12}, a_{13}, a_{21}, a_{22}, a_{23}, a_{31}, a_{32}, a_{33}$.

19. Which of the following matrices are scalar matrices?

(*i*) $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ (*ii*) $\begin{pmatrix} 5 & 1 \\ 0 & 5 \end{pmatrix}$ (*iii*) $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ (*iv*) $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

20. If $A = \begin{pmatrix} 1 & 3 & 4 \\ 1 & 2 & 3 \end{pmatrix}$, find a matrix B such that $A + B$ is a zero matrix.

21. If $A = \begin{pmatrix} 0 & 1 & 2 \\ 2 & 3 & 4 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{pmatrix}$, $C = \begin{pmatrix} 2 & 3 & 4 \\ 4 & 5 & 6 \end{pmatrix}$, then compute the following:

(*i*) $(A + B) + C$ (*ii*) $(A - B) + C$ (*iii*) $A - B - C$

(*iv*) $2A + 3B$ (*v*) $A + 2B + 3C$

NOTES

22. If $A = \begin{pmatrix} -2 & 3 & -1 \\ -1 & 2 & -1 \\ -6 & 9 & -4 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 3 & -1 \\ 2 & 2 & -1 \\ 3 & 0 & -1 \end{pmatrix}$, then show that $AB = BA$.

23. If $A = (1 \quad 2 \quad 3)$, $B = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$, then verify $(AB)' = B'A'$.

24. If $A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 1 & 3 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 3 & 4 \\ 5 & 6 & 7 \end{pmatrix}$, then verify $(A + B)' = A' + B'$.

25. If $A = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$, then verify $(A'B') = A$.

Long-Answer Questions

1. How many ATM cells can be transmitted in 15 seconds over a link operating at the following rates:
   (a) 1 megabit per second (1 megabit = 1,000,000 bits)
   (b) 128 KB per second
2. Let f is an invertible function from B to C and g is an invertible function from A to B. Prove that the inverse of the composition $(f \circ g)$ is given by $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$.
3. The first nine digits of the ISBN of the 2nd edition of *Discrete Mathematics* by C.L. Liu are 0-07-100544. What is the check digit for that book?
4. The ISBN of *Introduction to Languages and the Theory of Computation* (2nd edition) is 0-07-46 Q 722-3, where Q is a digit. Find the value of Q.
5. Find whether the check character of the ISBN for the book '*Systems Programming*' by Donovan was computed correctly by the publisher.
6. Which memory locations are assigned by the hashing function $h(k) = K \bmod 100$ to the records of students with the register numbers?
   (a) 10,457 (b) 100,321 (c) 560,076
7. A parking lot has 26 slots, numbered from 0 to 25, reserved for visitors. Visitors are assigned slots using the hashing function $h(k) = k \bmod 26$, where k is the number formed from the first two digits on a visitor's car number plate.
   (*i*) What slots are assigned by the hashing function to cars that have the following first two digits on their number plates?
   01, 11, 21, 22, 99.
   (*ii*) Describe a procedure that visitors should follow in a free parking space when collisions happen.

8. If x, y, z are in AP, then show that

   (*i*) $\frac{1}{yz}, \frac{1}{zx}, \frac{1}{xy}$ are in AP;

   (*ii*) $x^2(y + z), y^2(z + x), z^2(x + y)$ are in AP;

   (*iii*) $x\left(\frac{1}{y}+\frac{1}{z}\right), y\left(\frac{1}{z}+\frac{1}{x}\right), z\left(\frac{1}{x}+\frac{1}{y}\right)$ are in AP.

9. There are n arithmetic means between 3 and 54. If 8th Arithmetic Mean bears a ratio 3 : 5 to $(n – 2)$th arithmetic mean, find n.

10. The interior angles of a polygon of n sides are in AP, the smallest angle is of 42º and the common difference is 33º. Find n.

11. Between two numbers whose sum is $2\frac{1}{6}$, an even number of arithmetic means is inserted; the sum of these exceeds their number by unity. How many means are there?

12. If pth term of an AP is $1/q$ and qth term is $1/p$, show that the sum of pq terms is $\frac{1}{2}(pq + 1)$.

13. If p, q, r, s are any four consecutive terms of an AP, then show that $p^2 – 3q^2 + 3r^2 – s^2 = 0$

    (*Hint.* Let $p = \alpha – 3\beta$, $q = \alpha – \beta$, $r = \alpha + \beta$, $s = \alpha + 3\beta$)

14. A moneylender lends Rs 1000 and charges an overall interest of Rs 140. He recovers the loan and interest by 12 monthly instalments each less by Rs 10 than the preceding one. Find the amount of the first instalment.

15. To verify cash balances, the auditor of a certain bank employs his assistant to count cash in hand of Rs 4500. At first, he counts quietly at the rate of Rs 150 per minute for 10 minutes only but at the end of that time he begins to count at the rate of Rs 2 less every minute than he could count in the previous minute. Ascertain how much time he will take to count this sum of Rs 4500.

16. The sum of three numbers in a GP is 38 and their product is 1728. Find them.

17. There are n terms in GP, show that nth root of their product is equal to the square root of the product of its first and last terms.

18. The sum of an infinite number of terms of a GP is 4 and the sum of their cubes is 192. Find the series.

19. If the arithmetic mean between a and b $(a>b)$ is twice as great as the geometric mean, show that either $a/b = (2 + \sqrt{3})$ or $(2 – \sqrt{3})$.

NOTES

20. Find the sum of the series:

$$\frac{1}{2}+\frac{1}{3^2}+\frac{1}{2^3}+\frac{1}{3^4}+\frac{1}{2^5}+\frac{1}{3^6}+\cdots \text{ to } \infty$$

21. If $x = 1 + a + a^2 + \cdots + \infty$, $y = 1 + b + b^2 + \cdots + \infty$ prove that if $|a| < 1$ and $|b| < 1$

then $1 + ab + a^2b^2 + \cdots = \frac{xy}{x + y - 1}$

22. If a, b, c, d are in GP, then prove that

(*i*) $\frac{ab - cd}{b^2 - c^2} = \frac{a + c}{b}$

(*ii*) $(ab + bc + cd)^2 = (a^2 + b^2 + c^2)(b^2 + c^2 + d^2)$

23. Consider the matrices

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

(*a*) Show that $A^2 = B$ and $A^3 = I$. What is A^4?

(*b*) Show that $B^2 = I$ and $B^2 = A$.

(*c*) Show that $A^3 = BA = AB = B^3 = I$, hence A and B commute.

24. The production of a book involves several steps: first, it must be set in type, then it must be printed and, finally, it must be supplied with covers and bound. Suppose that typesetter charges Rs. 6 per hour, paper costs 1/4 paisa per sheet, the printer charges 11 paise for each minute that his press runs, the cover costs 28 paise and the binder charges 15 paise to bind each book. Suppose now that a publisher wishes to print a book that requires 300 hours of work by the typesetter, 220 sheets of paper per book and 5 minutes of press time per book.

(*i*) Using matrix multiplication, find the cost of publishing one copy of a book.

(*ii*) Using matrix addition and multiplication, find the cost of printing the first edition run of 5000 copies.

25. Suppose that a building contractor has accepted orders for five houses of type I, seven houses of type II and twelve houses of type III. The following matrix gives the amount of each raw material to be used in each type of house, expressed in convenient units:

$$R = \begin{array}{c} \\ \text{Type I} \\ \text{Type II} \\ \text{Type III} \end{array} \begin{pmatrix} \text{Steel} & \text{Wood} & \text{Glass} & \text{Paint} & \text{Labour} \\ 5 & 20 & 16 & 7 & 17 \\ 7 & 18 & 12 & 9 & 21 \\ 6 & 25 & 8 & 5 & 13 \end{pmatrix}$$

NOTES

Using matrix multiplication, find how much of each raw material the contractor should order.

26. Apply the elementary row operation $R_1 \leftrightarrow R_3$ in each of the following matrices:

(*i*) $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 6 & 7 & 8 & 9 \end{pmatrix}$ (*ii*) $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ (*iii*) $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$

27. Reduce each of the following matrices to the Echelon form:

(*i*) $\begin{pmatrix} 2 & 0 & -1 \\ 5 & 1 & 0 \\ 0 & 1 & 3 \end{pmatrix}$ (*ii*) $\begin{pmatrix} 1 & -1 & 1 \\ 4 & 1 & 0 \\ 8 & 1 & 1 \end{pmatrix}$

(*iii*) $\begin{pmatrix} -1 & -3 & 3 & -1 \\ 1 & 1 & -1 & 0 \\ 2 & -5 & 2 & -3 \\ -1 & 1 & 0 & 1 \end{pmatrix}$ (*iv*) $\begin{pmatrix} -1 & 0 & -2 & 2 \\ -2 & 1 & 0 & 1 \\ 1 & 0 & 2 & -1 \\ -4 & 1 & -3 & 1 \end{pmatrix}$

(*v*) $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \\ 4 & 5 & 6 \end{pmatrix}$ (*vi*) $\begin{pmatrix} 1 & -1 & 3 & 6 \\ 1 & 3 & -3 & -4 \\ 5 & 3 & 3 & 11 \end{pmatrix}$

(*vii*) $\begin{pmatrix} 2 & 3 & -1 & -1 \\ 1 & -1 & -2 & -4 \\ 3 & 1 & 3 & -2 \\ 6 & 3 & 0 & -7 \end{pmatrix}$ (*viii*) $\begin{pmatrix} 3 & 2 & -1 & 5 \\ 5 & 1 & 4 & -2 \\ 1 & -4 & 11 & -19 \end{pmatrix}$

(*ix*) $\begin{pmatrix} 1 & 2 & 1 \\ -1 & 0 & 2 \\ 2 & 1 & -3 \end{pmatrix}$ (*x*) $\begin{pmatrix} 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 3 & 4 \end{pmatrix}$

28. Solve the following equations:

$x + 2y + 3z = 14$

$3x + y + 2z = 11$

$2x + 3y + z = 11$

$x + y + z = 3$

$3x - 5y + 2z = 8$

$5x - 3y + 4z = 14$

29. Find the rank of the following matrices:

(*i*) $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 4 & 5 & 6 \end{pmatrix}$ (*ii*) $\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & -2 & 1 \\ 2 & 0 & -3 & 2 \\ 3 & 3 & 0 & 3 \end{pmatrix}$ (*iii*) $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \\ 8 & 13 & 21 \end{pmatrix}$

(*iv*) $\begin{pmatrix} 1 & 4 & 3 & 2 \\ 1 & 2 & 3 & 4 \\ 2 & 6 & 7 & 5 \end{pmatrix}$ (*v*) $\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & -2 & 1 \\ 2 & 0 & -3 & 2 \\ 3 & 3 & 0 & 3 \end{pmatrix}$

1.10 FURTHER READING

Anderson, David R., Dennis J. Sweeney and Thomas A. Williams. *Essentials of Statistics for Business and Economics*. Mumbai: Thomson Learning, 2007.

Anderson, David R., Dennis J. Sweeney and Thomas A. Williams. *Quantitative Methods for Business*. Mumbai: Thomson Learning, 2005.

Bhardwaj, R.S. *Business Statistics*. New Delhi: Excel Books, 2000.

Chandan, J.S. *Business Statistics.* New Delhi: Vikas Publishing House, 2004.

Gupta, C.B. and Vijay Gupta. *An Introduction to Statistical Methods*. New Delhi: Vikas Publishing House, 2004.

Hooda. R.P. *Statistics for Business & Economics.* New Delhi: Macmillan India Ltd., 2004.

Kothari C.R. *Quantitative Techniques*. New Delhi: Vikas Publishing House, 1984.

Levin, Richard I. and David S. Rubin. *Statistics for Business*. New Delhi: Prentice Hall of India, 1990.

Monga, G.S. *Mathematics and Statistics for Economics*. New Delhi: Vikas Publishing House.

Sancheti D.C. and V.K. Kapoor. *Business Mathematics*. New Delhi: Sultan Chand & Sons.

Zameeruddin Qazi, V.K. Sharma and S.K. Bhambri. *Business Mathematics*. New Delhi: Vikas Publishing House, 2008.

UNIT 2 FREQUENCY DISTRIBUTION AND SKEWNESS

NOTES

Structure

2.0 INTRODUCTION

In this unit, you will learn about frequency distribution and skewness. For a proper understanding of the quantitative data, they should be classified and converted into a frequency distribution. A frequency distribution is defined as the list of all the values obtained in the data and the corresponding frequency with which these values occur in the data. The various methods of measurement of the central tendency are mean, median and mode. You will learn the arithmetic procedures that can be used for analysing and interpreting quantitative data, i.e. the concept of arithmetic mean, median and mode, but classification is only the first step in statistical analysis.

This unit will also introduce you to data variations. Variance or coefficient of variation has the same properties as standard deviation and is the square of standard deviation, represented as σ^2.

NOTES

The unit talks about the measures of dispersion. The measures of central tendency are computed to see through the variability or *dispersion* of the individual values, but the dispersion is in itself a very important property of a distribution and needs to be measured by appropriate statistics.

You will also learn that measure of dispersion can be expressed in an 'absolute form', or in a 'relative form'. It is said to be in an absolute form when it states the actual amount by which the value of an item on an average deviates from a measure of central tendency.

2.1 UNIT OBJECTIVES

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

- Understand the concept of frequency distribution
- Understand the significance of central tendency in frequency distribution
- Calculate variance and coefficient of variation
- Understand the concept of measures of dispersion and its significance in statistical analysis
- Describe the various types of measures
- Describe the characteristics, merits and limitations of measures
- Understand when frequency distribution is called skewed

2.2 FREQUENCY DISTRIBUTION

Statistical data can be organized into a frequency distribution which simply lists the value of the variable and the frequency of its occurrence in a tabular form. A frequency distribution can then be defined as a list of all the values obtained in the data and the corresponding frequency with which these values occur in the data.

The frequency distribution can be either ungrouped or grouped. When the number of values of the variable is small, then we can construct an ungrouped frequency distribution, which is simply listing the frequency of occurrence against the value of the given variable. As an example, let us assume that 20 families were surveyed to find out the number of children in each family. The raw data obtained from the survey are as follows:

0, 2, 3, 1, 1, 3, 4, 2, 0, 3, 4, 2, 2, 1, 0, 4, 1, 2, 2, 3

These data can be classified into an ungrouped frequency distribution. The number of children becomes our variable (X) for which we can list the frequency of occurrence (f) in a tabular form as given in Table 2.1.

Table 2.1 *Data Classification*

Number of Children (X)	*Frequency (f)*
0	3
1	4
2	6
3	4
4	3
	Total = 20

Table 2.1 is also known as discrete frequency distribution, where the variable has discrete numerical values.

However, when the data set is very large, it becomes necessary to condense the data into a suitable number of groups or classes of the variable values and then assign the combined frequencies of these values into their respective classes. As an example, let us assume that 100 employees in a factory were surveyed to find out their ages. The youngest person was 20 years of age and the oldest was 50 years old. We can construct a grouped frequency distribution for these data so that instead of listing frequency by every year of age, we can list frequency according to an age group. Also, since age is a continuous variable, the frequency distribution would be as given in Table 2.2.

Table 2.2 *Frequency Distribution*

Age Group (Years)	*Frequency*
20 to less than 25	5
25 " " " 30	15
30 " " " 35	25
35 " " " 40	30
40 " " " 45	15
45 " " " 50	10
	Total = 100

In this example, all persons between 20 years (including 20 years old) and 25 years (but not including 25 years old) would be grouped in the first class, and so on. The interval of 20 to less than 25 is known as class interval (CI). A single representation of a class interval would be the midpoint (or average) of that class interval. The midpoint is also known as the class mark.

2.2.1 Constructing a Frequency Distribution

The number of groups and the size of class intervals are more or less arbitrary in nature within the general guidelines established for constructing a frequency distribution. The following guidelines for such a construction may be considered:

NOTES

(*i*) The classes should be clearly defined and each of the observations should be included in only one of the class intervals. This means that the intervals should be chosen in such a manner that one score cannot belong to more than one class interval, so that there is no overlapping of class intervals.

(*ii*) The number of classes should neither be too large nor too small. Normally, between six and fifteen classes are considered to be adequate. Fewer class intervals would mean a greater class interval width with consequent loss of accuracy. Too many class intervals result in greater complexity.

(*iii*) All intervals should be of the same width. This is preferred for easy computations. A suitable class width can be obtained by knowing the range of data (which is the absolute difference between the highest value and the lowest value in the data) and the number of classes which are predetermined, so that

$$\text{The width of the interval} = \frac{\text{Range}}{\text{Number of classes}}$$

In the case of ages of factory workers where the youngest worker was 20 years old and the oldest was 50 years old, the range would be 50 – 20 = 30. If we decide to make 10 groups, then the width of each class would be:

$30/10 = 3$

Similarly, if we decide to make 6 classes instead of 10, then the width of each class interval would be:

$30/6 = 5$

(*iv*) Open-ended cases—where there is no lower limit of the first group or no upper limit of the last group—should be avoided since this creates difficulty in analysis and interpretation. (The lower and the upper values of a class interval are known as lower and upper limits.)

(*v*) Intervals should be continuous throughout the distribution. For example, in the case of factory workers, we could put them in groups of 20 to 24 years, then 25 to 29 years, and so on, but it would be highly misleading because it does not accurately represent the person who is between 24 and 25 years or between 29 and 30 years, and so on. Accordingly, it is more representative to group them as 20 years to less than 25 years and 25 years to less than 30 years. In this way, everybody who is 20 years and a fraction less than 25 years is included in the first category and the person who is exactly 25 years and above but a fraction less than 30 years would be included in the second category, and so on. This is especially important for continuous distributions.

(*vi*) The lower limits of class intervals should be simple multiples of the interval width. This is primarily for the purpose of simplicity in construction and interpretation. In our example of 20 years but less than 25 years, 25 years

but less than 30 years, and 30 years but less than 35 years, the lower limit values for each class are simple multiples of the class width which is 5.

Example 2.1: The ages (in years) of a sample of 30 persons are as follows:

20, 18, 25, 68, 23, 25, 16, 22, 29, 37,
35, 49, 42, 65, 37, 42, 63, 65, 49, 42,
53, 48, 65, 72, 69, 57, 48, 39, 58, 67.

Construct a frequency distribution for these data.

Solution:

Follow the steps as given:

1. Find the range of the data by subtracting the lowest age from the highest age. The lowest value is 16 and the highest value is 72. Hence, the range is 72 – 16 = 56.
2. Assume that we shall have 6 classes, since the number of values is not too large. Now we divide the range of 56 by 6 to get the width of the class interval. The width is 56/6 = 9.33. For the sake of convenience, assume the width to be 10 and start the first class boundary with 15 so that the intervals would be 15 and upto 25, 25 and upto 35, and so on.
3. Combine all the frequencies that belong to each class interval and assign this total frequency to the corresponding class interval as follows:

Class Interval (years)	*Tally*	*Frequency (f)*
15 to less than 25	𝍸	5
25 to less than 35	𝍸	3
35 to less than 45	𝍸 ||	7
45 to less than 55	𝍸	5
55 to less than 65	𝍸	3
65 to less than 75	𝍸 ||	7
	Total = 30	

Conversion of a discrete list to a continuous list

In statistics, calculations are performed by arranging the large raw (ungrouped) data set into grouped data and are represented in tabular form called frequency distribution table. The data to be grouped must be homogenous and comparable. The frequency distribution table gives the size and the number of class intervals. The range of each class is defined by the class boundaries.

The variables constitute a discrete list or a continuous list. A variable is considered as continuous when it can assume an infinite number of real values and it

NOTES

is considered discrete when it is the finite number of real values. Examples of a continuous variable are distance, age, temperature and height measurements, whereas those of a discrete variable are the scores given by the experts or the judgement team for a competitive examination, basket ball match, cricket match, etc.

For a discrete list of data, the range can be defined as 0 – 4, 5 – 9, 10 – 14, and so on. Similarly, the range of data for a continuous list can be defined as 10 – 20, 20 – 30, 30 – 40, and so on.

In a class interval, the end points define the lowest and the highest values that a variable can take. In this example, if we consider the data set for age, then the class intervals are 0 to 4 years, 5 to 9 years, 10 to 14 years, and 14 years and above. For a discrete variable, the end points of the first class interval are 0 and 4, but for a continuous variable it will be 0 and 4.999. In this way, the discrete variables can be converted into continuous variables and vice versa.

Conversion of an ungrouped list into a grouped list

The data collected first-hand for any statistical evaluation are considered as raw or ungrouped data as these are not meaningful and do not present a clear picture. These are then arranged in the ascending or the descending order in a tabular form called array. Example 2.2 will make the concept more clear.

Example 2.2: The following table shows the daily wages (in Rs) of 40 workers. Convert the ungrouped data into grouped data and also prepare a discrete frequency table with tally marks.

Ungrouped Data

90	85	50	70	55	86	60	75	80	65
75	78	86	80	60	90	55	95	65	85
55	70	60	85	80	95	90	75	60	86
60	95	85	70	65	55	86	90	80	78

Solution: After arranging this into grouped data, we get the following table:

Grouped Data

95	95	95	90	90	90	90	86	86	86
86	85	85	85	85	80	80	80	80	78
78	75	75	75	70	70	70	65	65	65
60	60	60	60	60	55	55	55	55	50

NOTES

The discrete frequency distribution of daily wages with tally marks:

Daily Wages	*Tally Marks*	*Frequency*
95	|||	3
90	||||	4
86	||||	4
85	||||	4
80	||||	4
78	||	2
75	|||	3
70	|||	3
65	|||	3
60	𝍸	5
55	||||	4
50	|	1

Class intervals of unequal width

From the data given in Example 2.2, a table showing class intervals of unequal width is drawn.

Daily Wages	*Tally Marks*	*Frequency*
50 – 55	𝍸	5
55 – 60	𝍸	5
60 – 65	|||	3
65 – 70	|||	3
70 – 75	|||	3
75 – 78	||	2
78 – 80	||||	4
80 – 85	||||	4
85 – 86	||||	4
86 – 90	||||	4
90 – 95	|||	3
	Total	40

NOTES

Cumulative frequency distribution

While the frequency distribution table tells us the number of units in each class interval, it does not tell us directly the total number of units that lie below or above the specified values of class intervals. This can be determined from a cumulative frequency distribution. When the interest of the investigator focuses on the number of items below a specified value, then this specified value is the upper limit of the class interval. It is known as less than cumulative frequency distribution. Similarly, when the interest lies in finding the number of cases above a specified value, then this value is taken as the lower limit of the specified class interval and is known as more than cumulative frequency distribution. The cumulative frequency simply means summing up the consecutive frequencies as follows (taking Example 2.1):

Class Interval (Years)	*(f)*	*Cumulative Frequency (Less than)*
15 and up to 25	5	5 (less than 25)
25 and up to 35	3	8 (less than 35)
35 and up to 45	7	15 (less than 45)
45 and up to 55	5	20 (less than 55)
55 and up to 65	3	23 (less than 65)
65 and up to 75	7	30 (less than 75)

Similarly, the following is the greater than cumulative frequency distribution:

Class Interval (Years)	*(f)*	*Cumulative Frequency (Greater than)*
15 and up to 25	5	30 (greater than 15)
25 and up to 35	3	25 (greater than 25)
35 and up to 45	7	22 (greater than 35)
45 and up to 55	5	15 (greater than 45)
55 and up to 65	3	10 (greater than 55)
65 and up to 75	7	7 (greater than 65)

In this greater than cumulative frequency distribution, 30 persons are older than 15, 25 are older than 25, and so on.

2.2.2 Preparing a Frequency Distribution Table

If a frequency table records the distribution of a discrete variable, then the real and the apparent class limits are the same (unless the class interval is exclusive). This is due to the fact that discrete data are always expressed in whole numbers and are always characterized by gaps at which no measure may ever be found. Thus, if the class intervals of discrete variable are

6–10 11–15 16–20

the apparent limits 6 and 10, 11 and 15, 16 and 20 are real limits. The class 6–10 includes only those items whose sizes are 6, 7, 8, 9 or 10. Any item whose size is more than 10, i.e. 11, 12, etc., or less than 6, i.e. 5 is not included in this class, but in the next higher or the next lower class. There are, of course, no values between 10 and 11 or 15 and 16. In such a case, the midpoint is the middle of the five values included in a class, namely 8 is the midpoint in 6–10 class, 13 is the midpoint in 11–15 class, and 18 is the midpoint in 16–20 class.

If, however, the class interval is exclusive, the apparent limits are not real and before finding the midpoint, the real limits should be determined. If the class interval is given in the following manner, it is said to be an exclusive class interval:

(*i*) 5–10 (*ii*) 10–15 (*iii*) 15–20

This means that an item having a value 15 is to be included either in class (*ii*) or class (*iii*). If it is included in class (*ii*) it means value 10 is included in class (*i*). Hence, the real limits of class (*ii*) are 11–15 and the midpoint is 13. If 15 is not included in class (*ii*) but is included in class (*iii*) the real limits of class (*ii*) are then 10–14 and the midpoint is 12. It, therefore, follows that whenever we have an exclusive class interval, we must decide as to which limit of the class is excluded and only then the midpoint should be ascertained.

If the frequency table records the distribution of a continuous variable, then the real limits are not the same as the apparent limits. This is because theoretically such variables can be measured to an infinitesimal fraction of a unit, and the measures that are obtained are only approximations to absolute accuracy. While measuring the weight of boys, for example, we seldom go to a unit smaller than the pound. Thus, when we say that the weight of an individual is 140 pounds, what we really mean is that his weight is nearer 140 pounds than 139 or 141 pounds. This means that it is somewhere between 139.5 and 140.5 pounds.

From this, it follows that if in any frequency distribution of weights we find a class interval identified by the interval limit (say, 140–144), we must conclude that (*a*) weights have been measured correct to the nearest pound and (*b*) hence, the real limits of the interval extend by 0.5 pounds on either side and class interval, strictly speaking, is 139.5–144.5. The midpoint of this class is to be determined from these limits. The method of finding the mid-value in this case is as follows:

$$\text{Lower limit of the class} + \frac{\text{Upper limit} - \text{Lower limit}}{2} = 139.5 + \frac{5}{2} = 139.5 + 2.5 = 142.0$$

or $$\frac{\text{Upper limit} + \text{Lower limit}}{2} = \frac{139.5 + 144.5}{2} = \frac{284}{2} = 142.0$$

If the weight has been measured correct to the nearest tenth of a pound, we will have class intervals like the following:

140 – 144.9

145 – 149.9

NOTES

On the basis of what has been said earlier, the real limits are:

139.95 – 144.95

144.95 – 149.95

Here the midpoint will be, $\frac{139.95 + 144.95}{2}$ = 142.45, i.e. 142.5

Example 2.3: The first and third columns of the following table give the frequency distribution of the average monthly earnings of male workers. Calculate the mean earnings.

Distribution of Male Workers by Average Monthly Earnings (Computation of Arithmetic Mean by Long Method)

Monthly Earnings Rs	*Midpoint Rs (m)*	*No. of Workers (f)*	*Rs f(m)*
27.5–32.5	30	120	3,600
32.5–37.5	35	152	5,320
37.5–42.5	40	170	6,800
42.5–47.5	45	214	9,630
47.5–52.5	50	410	20,500
52.5–57.5	55	429	23,595
57.5–62.5	60	568	34,080
62.5–67.5	65	650	42,250
67.5–72.7	70	795	55,650
72.5–77.5	75	915	68,625
77.5–82.5	80	745	59,600
82.5–87.5	85	530	45,050
87.5–92.5	90	259	23,310
92.5–97.5	95	152	14,440
97.5–102.5	100	107	10,700
102.5–107.5	105	50	5,250
107.5–112.5	110	25	2,750
		Σf = 6,291	$\Sigma f(m)$ = 4,31,150

Solution: Since the variable is a continuous one, the midpoints are calculated simply as (lower limit + upper limit)/2, and are shown in second column as *m*. The *f(m)* values are calculated in column 4. The mean is calculated as

$$\bar{X} = \frac{\Sigma f(m)}{\Sigma f} = \frac{4,31,150}{6,291} = \text{Rs } 68.5$$

If we compute the arithmetic mean from unclassified data, it may differ slightly from Rs 68.5. This lack of agreement is due to the inadequacy of the mid-value assumption. It is true that none of the mid-value is actually the true concentration point of this class. But in the case of symmetrical distributions, there is greater possibility of errors compensating some of the midpoints erring by being too low and others erring by being too high. However, if the frequency tails off towards

either the high or low values, i.e. if it departs seriously from a symmetrical distribution, the arithmetic average computed will be somewhat in error because of the failure of the known errors in the midpoints assumption to compensate.

NOTES

CHECK YOUR PROGRESS

1. Differentiate a continuous variable from a discrete variable.
2. What is the less than cumulative frequency distribution?

2.3 FREQUENCY DISTRIBUTION AND MEASURES OF CENTRAL TENDENCY

2.3.1 Descriptive Statistics

A single number describing some feature of a frequency distribution is called *descriptive statistics*. The main thrust of a statistician presenting a mass of data is to evolve few such descriptive statistics which describe the essential nature of the frequency distribution.

For a proper appreciation of the various descriptive statistics involved, it is necessary to note that most of the statistical distributions have some common features. Though the size of the variables varies from item to item, most of the items are distributed in such a manner that if you move from the lowest value to the highest value of the variable, the number of items at each successive stage increases with a certain amount of regularity till you reach a maximum; and then, as you proceed further, it decreases with the similar regularity. If you plot the percentage frequency density, i.e. the percentage of cases in an interval of *unit variable width*, you get frequency curves of the type shown in Figure 2.1. (Note that the area under each curve should be equal to 100, the total percentage points.)

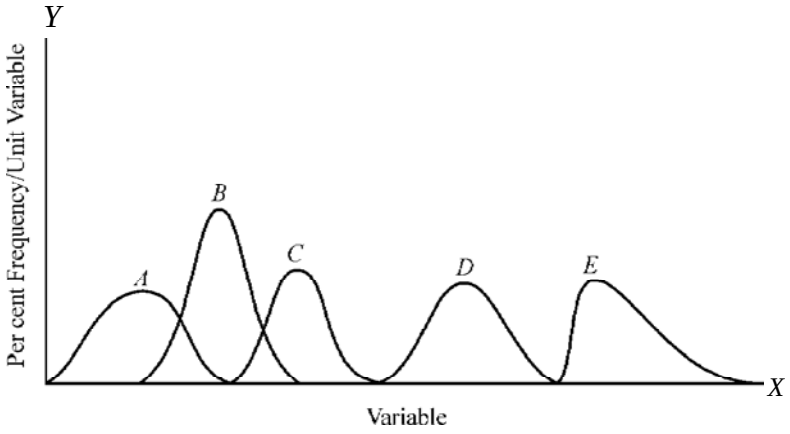


Fig. 2.1 *Frequency Curves*

There are various 'gross' ways in which frequency curves can differ from one another. Even when the 'general' shapes of the curves are the same (the area

under them already made equal by the strategy of plotting the per cent density), the details of the shape may change. Thus, the curve *B* has a smaller spread than *A*, the curve *C* is more peaky and curve *E* is less symmetrical. Even when the curves have almost the same shape (i.e. same spread, peakiness, symmetry, etc.) as in curves *A* and *D*, the two may differ in location along the variable axis. Thus, the items of distribution *D* are generally larger than those of *A*. So also are those of *B* compared to *A*. Hence, a kind of an 'average' location of the distribution along the variable axis is an important descriptive statistics. These statistics are collectively known as measures of location or of central tendency.

2.3.2 Measures of Central Tendency

As already mentioned, these statistics indicate the location of the frequency curve along the *X*-axis and ignore all other features of the distribution. There are various possible measures that can be used to 'locate' a frequency distribution, as shown in Figure 2.2.

A, the minimum value.

B, the value of maximum concentration.

C, the value which divides the distribution into half, such that one half of the items have value less than this and the other half more.

D, the average value of all items.

E, the 95th percentile, i.e. the value below which 95 per cent items lie.

F, the maximum value.

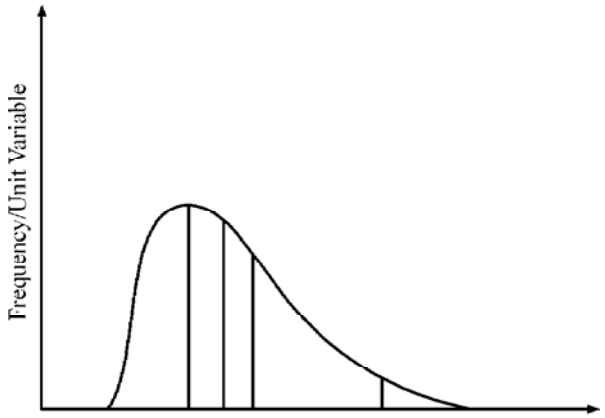


Fig. 2.2 *Frequency Distribution*

If the shape of the frequency distributions were fixed, then all these measures are equally descriptive, and fix the location of the curve. But, the practical distributions that we deal with always have some change in shape depending on the samples we take, even though the general shapes are quite similar. It is, therefore, necessary that we choose those measures of location which are not very sensitive to the specific values of items, in particular the extreme values. Thus, measures *A*

NOTES

and *E* are generally meaningless because they depend on the values of the lowest and the highest items, respectively. The other measures, on the contrary, are less susceptible to extreme values because they are somehow related to the entire distributions. Thus, we treat *B*, *C*, *D* and *E* as the most common measures of location. There are some more of such measures which we will consider later.

The most important object of calculating and measuring central tendency is to determine a 'single figure' which may be used to represent a whole series involving magnitudes of the same variable. In that sense, it is an even more compact description of the statistical data than the frequency distribution.

Since an 'average' represent the entire data, it facilitates comparison within one group or between groups of data. Thus, the performance of the members of a group can be compared by relating it to the average performance of the group. Likewise, the achievements of groups can be compared by a comparison of their respective averages.

2.3.3 Arithmetic Mean

There are several commonly used measures such as arithmetic mean, mode and median. These values are very useful not only in presenting the overall picture of the entire data but also for the purpose of making comparisons among two or more sets of data.

While arithmetic mean is the most commonly used measure of central location, mode and median are more suitable measures under certain sets of conditions and for certain types of data. However, each measure of central tendency should meet the following requisites.

1. It should be easy to calculate and understand.
2. It should be rigidly defined. It should have only one interpretation so that the personal prejudice or bias of the investigator does not affect its usefulness.
3. It should be representative of the data. If it is calculated from a sample, then the sample should be random enough to be accurately representing the population.
4. It should have sampling stability. It should not be affected by sampling fluctuations. This means that if we pick 10 different groups of college students at random and compute the average of each group, then we should expect to get approximately the same value from each of these groups.
5. It should not be affected much by extreme values. If few very small or very large items are present in the data, they will unduly influence the value of the average by shifting it to one side or other, so that the average would not be really typical of the entire series. Hence, the average chosen should be such that it is not unduly affected by such extreme values.

NOTES

Let us consider the measure of central tendency, arithmetic mean.

This is also commonly known as simply the mean. Even though average, in general, means any measure of central location. When you use the word average in you daily routine, you always mean the arithmetic average. The term is widely used by almost every one in daily communication. You speak of an individual being an average student or of average intelligence. You always talk about average family size or average family income or grade point average (GPA) for students, and so on.

For discussion purposes, let us assume a variable X which stands for some scores such as the ages of students. Let the ages of 5 students be 19, 20, 22, 22 and 17 years. Then variable X would represent these ages as follows:

X: 19, 20, 22, 22, 17

Placing the Greek symbol Σ(Sigma) before X would indicate a command that all values of X are to be added together. Thus,

$$\Sigma X = 19 + 20 + 22 + 22 + 17$$

The mean is computed by adding all the data values and dividing it by the number of such values. The symbol used for sample average is $\overline{X}$ so that

$$\overline{X} = \frac{19 + 20 + 22 + 22 + 17}{5}$$

In general, if there are n values in the sample, then

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

In other words,

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \qquad i = 1,\ 2,\ ...,\ n \tag{2.1}$$

Formula (2.1) states that add up all the values of X_i where the value of i starts at 1 and ends at n with unit increments so that $i = 1, 2, 3, ..., n$.

If instead of taking a sample, you take the entire population in your calculations of the mean, then the symbol for the mean of the population is μ (mu) and the size of the population is N, so that

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}, \qquad i = 1,\ 2,\ ...,\ N \tag{2.2}$$

NOTES

If you have the data in grouped discrete form with frequencies, then the sample mean is given by

$$\overline{X} = \frac{\Sigma f(X)}{\Sigma f} \qquad (2.3)$$

where Σf = Summation of all frequencies
= n

$\Sigma f(X)$ = Summation of each value of X multiplied by its corresponding frequency (f)

Example 2.4: Let us take the ages of 10 students as follows:

19, 20, 22, 22, 17, 22, 20, 23, 17, 18

Solution: These data can be arranged in a frequency distribution as follows:

Age *(X)*	*Frequency* *(f)*	*f(X)*
17	2	34
18	1	18
19	1	19
20	2	40
22	3	66
23	1	23
Total = 10		200

In the given case, you have $\Sigma f = 10$ and $\Sigma f(X) = 200$, so that

$$\overline{X} = \frac{\Sigma f(X)}{\Sigma f}$$

$$= 200/10 = 20$$

Example 2.5: Calculate the mean of the marks of 46 students given in the following table.

Frequency of Marks of 46 Students

Marks (*X*)	*Frequency* (*f*)
9	1
10	2
11	3
12	6
13	10
14	11
15	7
16	3
17	2
18	1
Total	46

NOTES

Solution: This is a discrete frequency distribution, and is calculated using equation (3.3). The following table shows the method of obtianing $\Sigma f(X)$.

Marks (X)	*Frequency (f)*	*f(X)*
9	1	9
10	2	20
11	3	33
12	6	72
13	10	130
14	11	154
15	7	105
16	3	48
17	2	34
18	1	18
	$\Sigma f = 46$	$\Sigma f(X) = 623$

Using equation (2.3), we get

$$\bar{X} = \frac{\Sigma f(X)}{\Sigma f} = \frac{623}{46} = 13.54$$

Arithmetic mean of grouped data

If, however, the data is grouped such that you are given frequency of finite-sized class intervals you do not know the value of every item. The calculation of arithmetic mean in such a case is then necessarily a process of estimation, based on some assumption. The standard assumption for this purpose is that all the items within a particular class are concentrated at the mid-value of the class and thus *f(X)* corresponding to the *f* items of a class equals *f(m)*, where *m* is the midpoint of the class interval, and the arithmetic mean is then given by

$$\bar{X} = \frac{\Sigma f(m)}{\Sigma f} \qquad (2.4)$$

The determination of the midpoint of a class interval requires some consideration. The position of the midpoint is determined by *real* as distinguished from *apparent class* limits.

Advantages of arithmetic mean

The advantages of arithmetic mean are as follows:

1. Its concept is familiar to most people and is intuitively clear.
2. In it, every data set has a mean, which is unique and describes the entire data to some degree. For example, when you say that the average salary of a professor is Rs 25,000 per month, it gives you a reasonable idea about the salaries of professors.
3. It is a measure that can be easily calculated.
4. It includes all values of the data set in its calculation.
5. Its value varies very little from sample to sample taken from the same population.

6. It is useful for performing statistical procedures such as computing and comparing the means of several data sets.

Disadvantages of arithmetic mean

NOTES

The disadvantages of arithmetic mean are as follows:

1. It is affected by extreme values, and hence, is not very reliable when the data set has extreme values, especially on one side of the ordered data. Thus, a mean of such data is not truly a representative of the data. For example, the average age of three persons of ages 4, 6 and 80 years gives you an average of 30.
2. It is tedious to compute for a large data set as every point in the data set is to be used in computations.
3. It does not allow to compute the mean for a data set that has open-ended classes either at the high or at the low end of the scale.
4. It cannot be calculated for qualitative characteristics, such as beauty or intelligence, unless these can be converted into quantitative figures, such as intelligence into IQs.

Properties of arithmetic mean

The arithmetic mean has the following interesting properties.

1. The sum of the deviations of individual values of X from the mean will always add up to zero. This means that if you subtract all the individual values from their mean, then some values will be negative and some will be positive, but if all these differences are added together, then the total sum will be zero. In other words, the positive deviations must balance the negative deviations, or symbolically:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0, \; i = 1, 2, ..., n$$

2. The second important characteristic of the mean is that it is very sensitive to extreme values. Since the computation of the mean is based upon inclusion of all values in the data, an extreme value in the data would shift the mean towards it, thereby making the mean unrepresentative of the data.
3. The third property of the mean is that the sum of squares of the deviations about the mean is minimum. This means that if you take differences between individual values and the mean and square these differences individually and then add these squared differences, then the final figure will be less than the sum of the squared deviations around any other number other than the mean. Symbolically, it means that

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \text{Minimum}, \; i = 1, 2, ..., n$$

NOTES

4. The product of the arithmetic mean and the number of values on which the mean is based is equal to the sum of all given values. In other words, if you replace each item in series by the mean, then the sum of the these substitutions will equal the sum of individual items. Thus, in the figures 3, 5, 7, 9, if we substitute the mean for each item, 6, 6, 6, 6 then the total is 24, both in the original series and in the substitution series.

   This can be shown as

   Since, $$\bar{X} = \frac{\Sigma X}{N}$$

   $\therefore$ $$N\bar{X} = \Sigma X$$

   For example, you we have a series of values 3, 5, 7, 9, the mean is 6. The squared deviations are:

X	$X - \bar{X} = X'$	X'^2
3	$3-6=-3$	9
5	$5-6=-1$	1
7	$7-6=1$	1
9	$9-6=3$	9
		$\Sigma X'^2 = 20$

   This property provides a test to check if the computed value is the correct arithmetic mean.

Example 2.6: The mean age of a group of 100 persons (grouped in intervals 10–, 12–, ..., etc.) was found to be 32.02. Later, it was discovered that age 57 was misread as 27. Find the corrected mean.

Solution: Let the mean be denoted by X. So putting the given values in the formula of A.M., you have,

$$32.02 = \frac{\sum X}{100}, \text{ i.e. } \sum X = 3202$$

$$\text{Correct } \sum X = 3202 - 27 + 57 = 3232$$

$\therefore$ $$\text{Correct AM} = \frac{3232}{100} = 32.32$$

Example 2.7: The mean monthly salary paid to all employees in a company is Rs 500. The monthly salaries paid to male and female employees average Rs 520 and Rs 420, respectively. Determine the percentage of males and females employed by the company.

Solution: Let N_1 be the number of males and N_2 be the number of females employed by the company. Also let x_1 and x_2 be the monthly average salaries paid to male and female employees and x_{12} be the mean monthly salary paid to all the employees.

$$\bar{x}_{12} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2}$$

or $$500 = \frac{520N_1 + 420N_2}{N_1 + N_2} \quad \text{or} \quad 20N_1 = 80N_2$$

or $$\frac{N_1}{N_2} = \frac{80}{20} = \frac{4}{1}$$

NOTES

Hence, the males and females are in the ratio of 4 : 1 or 80 per cent are males and 20 per cent are females in those employed by the company.

Short-cut methods for calculating mean

You can simplify the calculations of mean by noticing that if you subtract a constant amount A from each item X to define a new variable $X' = X - A$, the mean $\bar{X}'$ of X' differs from $\bar{X}$ by A. This generally simplifies the calculations and you can then add back the constant A, termed as the *assumed mean*[1]:

$$\bar{X} = A + \bar{X}' = A + \frac{\sum f(X')}{\sum f}$$

The following table illustrates the procedure of calculation by short-cut method using the data given in example 2.4. The choice of A is made in such a manner as to simplify calculation the most, and is generally in the region of the concentration of data.

X	*(f)*	*Deviation from Assumed Mean (13) X'*	*f(X')*
9	1	–4	–4
10	2	–3	–6
11	3	–2	–6
12	6	–1	–6
13	10	0	–22[1]
14	11	+1	+11
15	7	+2	+14
16	3	+3	+9
17	2	+4	+8
18	1	+5	+ 5
			+47
			–22
	$\Sigma f = 46$		$\Sigma fX' = 25$

The mean,

$$\bar{X} = A + \frac{\sum f(X')}{\sum f} = 13 + \frac{25}{46} = 13.54$$

which is the same as calculated in Example 2.5.

[1] Since there will not be an entry in the *f(X')* column corresponding to $X' = 0$, we write the sum –22 of the negative entries in the *f(X')* column. The sum of the positive products in the *f(X')* column, i.e. 47, is also written as the total *N*. The final sum 25 is then easily obtained.

NOTES

In the case of grouped frequency data, the variable X is replaced by mid-value m, and in the short-cut technique, you subtract a constant value A from each m, so that the formula becomes:

$$\bar{X} = A + \frac{\sum f\,(m - A)}{\sum f}$$

In the cases where the *class intervals are equal*, you may further simplify calculatation by taking the factor i from the variable $m - A$ defining,

$$X' = \frac{m - A}{i}$$

Where i is the class width. It can be verified that when X' is defined, the mean of the distribution is given by

$$\bar{X} = A + \frac{\sum f\,(X')}{\sum f} \times i$$

The following examples will illustrate the use of short-cut method.

Example 2.8: The ages of twenty husbands and wives are given in the following table. Form a two-way frequency table showing the relationship between the ages of husbands and wives with class intervals 20–24; 25–29, etc.

Calculate the arithmetic mean of the two groups after the classification.

S.No.	*Age of Husband*	*Age of Wife*
1	28	23
2	37	30
3	42	40
4	25	26
5	29	25
6	47	41
7	37	35
8	35	25
9	23	21
10	41	38
11	27	24
12	39	34
13	23	20
14	33	31
15	36	29
16	32	35
17	22	23
18	29	27
19	38	34
20	48	47

NOTES

Solution:

Frequency Distribution of Age of Husbands and Wives

Age of Husband	*Age of wife*						
	20–24	*25–29*	*30–34*	*35–39*	*40–44*	*45–49*	*Total*
20–24	III						3
25–29	II	III					5
30–34			I	I			2
35–39		II	III	I			6
40–44				I	I		2
45–49					I	I	2
Total	5	5	4	3	2	1	20

Calculation of Arithmetic Mean of Husbands' Age

Class Intervals	*Mid-values* m	*Husband Frequency* (f_1)	$x_2' = \frac{m-37}{5}$	f_1x_1'
20–24	22	3	–3	–9
25–29	27	5	–2	–10
30–34	32	2	–1	–2
				– 21
35–39	37	6	0	0
40–44	42	2	1	2
45–49	47	2	2	4
				6
		$\Sigma f_1 = 20$		$\Sigma f_1x_1' = -15$

Arithmetic mean of husband's ages:

$$\bar{x} = \frac{\sum f_1x_1'}{N} \times i + A = \frac{-15}{20} \times 5 + 37 = 33.25$$

Calculation of Arithmetic Mean of Wives' Age

Class Intervals	*Mid-values* m	*Wife Frequency* (f_2)	$x_2' = \frac{m-37}{5}$	f_2x_2'
20–24	22	3	–3	–9
25–29	27	5	–2	–10
30–34	32	2	–1	–2
35–39	37	6	0	0
40–44	42	2	1	2
45–49	47	1	2	2
		$\Sigma f_2 = 19$		$\Sigma f_2x_2' = -17$

Arithmetic mean of wives ages:

$$\bar{x} = \frac{\sum f_2x_2'}{N} \times i + A = \frac{-17}{19} \times 5 + 37 = 41.47$$

The weighted arithmetic mean

In the computation of arithmetic mean, you need to give equal importance to each observation in the series. This equal importance may be misleading if the individual values constituting the series have different importance as in the following example:

The Raja Toy shop sells	
Toy cars at	Rs 3 each
Toy locomotives at	Rs 5 each
Toy aeroplanes at	Rs 7 each
Toy double-decker at	Rs 9 each

What shall be the average price of the toys sold, if the shop sells 4 toys, one of each kind?

Mean price, i.e. $\bar{x} = \frac{\sum x}{4} = \text{Rs}\,\frac{24}{4} = \text{Rs } 6$

In this case the importance of each observation (price quotation) is equal in as much as one toy of each variety has been sold. In the above computation of the arithmetic mean this fact has been taken care of by including 'once only' the price of each toy.

But if the shop sells 100 toys: 50 cars, 25 locomotives, 15 aeroplanes and 10 double deckers, the importance of the four price quotations to the dealer is **not equal** as a source of earning revenue. In fact their respective importance is equal to the number of units of each toy sold, i.e.

The importance of toy car	50
The importance of locomotive	25
The importance of aeroplane	15
The importance of double-decker	10

It may be noted that 50, 25, 15, 10 are the quantities of the various classes of toys sold. It is for these quantities that the term 'weights' is used in statistical language. Weight is represented by symbol 'w', and Σw represents the sum of weights.

While determining the 'average price of toy sold', these weights are of great importance and are taken into account in the manner illustrated below:

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4}{w_1 + w_2 + w_3 + w_4} = \frac{\sum wx}{\sum w}$$

When w_1, w_2, w_3, w_4 are the respective weights of x_1, x_2, x_3, x_4 which in turn represent the price of four varieties of toys, namely car, locomotive, aeroplane and double-decker, respectively.

$$\bar{x} = \frac{(50\times3)+(25\times5)+(15\times7)+(10\times9)}{50+25+15+10}$$

$$= \frac{(150)+(125)+(105)+(90)}{100} = \frac{470}{100} = \text{Rs } 4.70$$

NOTES

The following table summarizes the steps taken in the computation of the weighted arithmetic mean.

Weighted Arithmetic Mean of Toys Sold by the Raja Toy Shop

Toys	*Price per Toy* Rs *x*	*Number Sold* *w*	*Price × Weight* *xw*
Car	3	50	150
Locomotive	5	25	125
Aeroplane	7	15	105
Double Decker	9	10	90
		$\Sigma w = 100$	$\Sigma xw = 470$

$$\Sigma w = 100; \quad \Sigma wx = 470$$

$$\bar{x} = \frac{\Sigma wx}{\Sigma w} = \frac{470}{100} = 4.70$$

The weighted arithmetic mean is particularly useful where you have to compute the *mean of means.* If you are given two arithmetic means, one for each of two different series, in respect of the *same variable,* and are required to find the arithmetic mean of the combined series, the weighted arithmetic mean is the only suitable method of its determination.

Example 2.9: The arithmetic mean of daily wages of two manufacturing concerns A Ltd. and B Ltd. is Rs 5 and Rs 7, respectively. Determine the average daily wages of both concerns if the numbers of workers employed were 2000 and 4000, respectively.

Solution: (*i*) Multiply each average (namely 5 and 7) by the number of workers in the concern it represents.

(*ii*) Add up the two products obtained in (*i*).

(*iii*) Divide the total obtained in (*ii*) by the total number of workers.

Weighted Mean of Mean Wages of A Ltd. and B Ltd.

Manufacturing Concern	*Mean Wages* *x*	*Workers Employed* *w*	*Mean Wages × Workers Employed* *wx*
A Ltd.	5	2000	10,000
B Ltd.	7	4000	28,000
		$\Sigma w = 6{,}000$	$\Sigma wx = 38{,}000$

$$\bar{x} = \frac{\Sigma wx}{\Sigma w}$$

$$= \frac{38{,}000}{6{,}000}$$

$$= \text{Rs } 6.33$$

The preceding examples explain that 'arithmetic means and percentage' are not original data. They are derived figures and their importance is relative to the original data from which they are obtained. This relative importance must be taken into account by weighting while averaging them (means and percentage).

2.3.4 Median

The second measure of central tendency that has a wide usage in statistical works is the median. Median is that *value* of a variable which divides the series in such a manner that the number of items below it is equal to the number of items above it. In other words, half the total number of observations lie below the median, and half above it. The median is thus a positional average.

The median of ungrouped data is found easily if the items are first arranged in order of magnitude. The median may then be located simply by counting, and its value can be obtained by reading the value of the middle observations. If you have five observations whose values are 8, 10, 1, 3 and 5, the values are first arrayed: 1, 3, 5, 8 and 10. It is now apparent that the value of the median is 5, since two observations are below that value and two observations are above it. When there is an even number of cases, there is no actual middle item and the median is taken to be the average of the values of the items lying on either side of $(N + 1)/2$, where N is the total number of items. Thus, if the values of six items of a series are 1, 2, 3, 5, 8 and 10. The median is the value of item number $(6 + 1)/2 = 3.5$, which is approximated as the average of the third and the fourth items, i.e. (3+5)/2=4.

Thus, the steps required for obtaining median are:

1. Arrange the data as an array of increasing magnitude.

2. Obtain the value of the $(N+ 1)/2$th item.

Even in the case of grouped data, the procedure for obtaining median is straightforward as long as the variable is discrete or non-continuous as is clear from the following example.

Example 2.10: Obtain the median size of shoes sold from the following data.

Number of Shoes Sold by Size in One Year

Size	*Number of Pairs*	*Cumulative Total*
5	30	30
$5\frac{1}{2}$	40	70
6	50	120
$6\frac{1}{2}$	150	270
7	300	570
$7\frac{1}{2}$	600	1170
8	950	2120
$8\frac{1}{2}$	820	2940
9	750	3690

$9\frac{1}{2}$	440	4130
10	250	4380
$10\frac{1}{2}$	150	4530
11	40	4570
$11\frac{1}{2}$	39	4609
	Total	4609

Solution: Median is the value of $\frac{(N+1)}{2}$th $= \frac{4609+1}{2}$th $= 2305$th item. Since the items are already arranged in ascending order (sizewise), the size of 2305th item is easily determined by constructing the cumulative frequency. Thus, the median size of shoes sold is 81, the size of 2305th item.

In the case of grouped data with continuous variable, the determination of median is a bit more involved. Consider an example: the data relating to the distribution of male workers by average monthly earnings are given in the following table. Clearly, the median of 6291 cases is the earnings of (6291 + l)/2 = 3l46th worker arranged in ascending order of earnings.

From the cumulative frequency, it is clear that this worker has his income in the class interval 67.5–72.5. But it is impossible to determine his exact income. You, therefore, resort to approximation by assuming that the 795 workers of this class are distributed *uniformly* across the interval 67.5 – 72.5. The median worker is (3146–2713) = 433rd of these 795, and hence, the value corresponding to him can be approximated as

$$67.5 + \frac{433}{795} \times (72.5 - 67.5) = 67.5 + 2.73 = 70.23$$

Distribution of Male Workers by Average Monthly Earnings

Group No.	*Monthly Earnings (Rs)*	*No. of Workers*	*Cumulative No. of Workers*
1	27.5–32.5	120	120
2	32.5–37.5	152	272
3	37.5–42.5	170	442
4	42.5–47.5	214	656
5	47.5–52.5	410	1066
6	52.5–57.5	429	1495
7	57.5–62.5	568	2063
8	62.5–67.5	650	2713
9	67.5–72.5	795	3508
10	72.5–77.5	915	4423
11	77.5–82.5	745	5168
12	82.5–87.5	530	5698
13	87.5–92.5	259	5957
14	92.5–97.5	152	6109
15	97.5–102.5	107	6216
16	102.5–107.5	50	6266
17	107.5–112.5	25	6291
		Total	6291

NOTES

The value of the median can thus be put in the form of the formula:

$$Me = l + \frac{\frac{N+1}{2} - C}{f} \times i$$

where l is the lower limit of the median class, i is its width, f is its frequency, C the cumulative frequency up to (but not including) the median class and N is the total number of cases.

Location of median by graphical analysis

The median can quite conveniently be determined by reference to the ogive which plots the cumulative frequency against the variable. The value of the item below which half the items lie can easily be read from the ogive.

Example 2.11: Obtain the median of data given in the following table.

Monthly Earnings	*Frequency (f)*	*Less than*	*More than (Greater than)*
27.5	—	0	6291
32.5	120	120	6171
37.5	152	272	6019
42.5	170	442	5849
47.5	214	656	5635
52.5	410	1066	5225
57.5	429	1495	4796
62.5	568	2063	4228
67.5	650	2713	3578
72.5	795	3508	2783
77.5	915	4423	1868
82.5	745	5168	1123
87.5	530	5698	593
92.5	259	5957	334
97.5	152	6109	182
102.5	107	6216	65
107.5	50	6266	25
112.5	25	6291	0

Solution: It is clear that this is grouped data. The first class is 27.5–32.5, whose frequency is 120, and the last class is 107.5–112.5, whose frequency is 25.

The median can also be determined by plotting both 'less than' and 'more than' cumulative frequency as shown in Figure 2.3. It is obvious that the two curves should intersect at the median of the data.

Quarticles, deciles and percentiles

You know that the median is the value of the item which is located at the centre of the array. You can define other measures which are located at other specified points. For example, the Nth *percentile* of an array is the value of the item such that N per cent items lie *below* it. Clearly then, the Nth percentile P_n of grouped data is given by

$$P_n = l + \frac{\frac{nN}{100} - C}{f} \times i$$

NOTES

where l is the lower limit of the class in which nN/100th item lies, i its width, f its frequency, C the cumulative frequency up to (but not including) this class, and N is the total number of items.

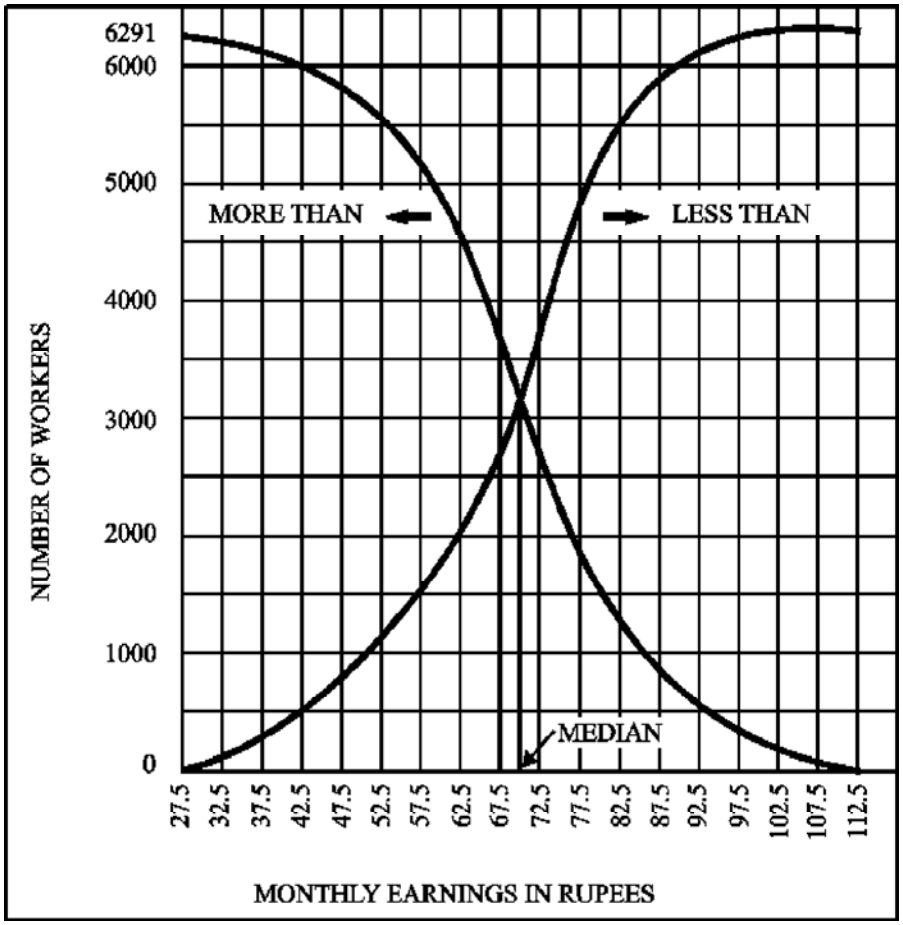


Fig. 2.3 *Location of Median*

You can similarly define the Nth *decile* as the value of the item below which (nN/10) items of the array lie. Clearly,

$$D_n = P_{10n} = l + \frac{\frac{nN}{10} - C}{f} \times i$$

The other most commonly referred to measures of location are the quartiles. nth quartile is the value of the item which lies at the $n(N/5)$th item. Clearly Q_2, the second quartile is the median. For grouped data,

$$Q_n = P_{25n} = l + \frac{\frac{nN}{4} - C}{f} \times i$$

NOTES

Example 2.12: Find the first and the third quartiles and the 90th percentile of the data given in the table of solution of Example 2.10.

Solution: The first quartile Q_1 is the value of the $N/4 = 6291/4 = 1572.75$th item. Thus, the appropriate class is 57.5–62.5, and by the preceding equation of decile,

$$Q_1 = 57.5 + \frac{(1572.75 - 1495)}{568} \times 5 = 58.18$$

The third quartile Q_3 is the value of the $3N/4 = 3 \times 6291 = 4718.25$th item, or Q_3 class interval is 77.5–82.5. Thus,

$$Q_3 = 77.5 + \frac{(4718.25 - 4423)}{745} \times 5 = 79.5$$

Similarly, P_{90} lies in 82.5–87.5 class interval, and

$$P_{90} = 82.5 + \frac{(5661.9 - 5168)}{530} \times 5 = 87.16$$

or 90 per cent workers earn less than Rs 87.16.

2.3.5 Mode

The mode is that value of the variable which occurs or repeats itself the greatest number of times. The mode is the most 'fashionable' size in the sense that it is the most common and typical, and is defined by Zizek as 'the value occurring most frequently in a series (or group of items) and around which the other items are distributed most densely.'

The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It is the most frequent or the most common value, provided that a sufficiently large number of items are available to give a smooth distribution. It will correspond to the value of the maximum point (ordinate) of a frequency distribution if it is an 'ideal' or smooth distribution. It may be regarded as the most typical of a series of values. The modal wage, for example, is the wage received by more individuals than any other wage. The model 'hat' size is that which is worn by more persons than any other single size.

It may be noted that the occurrence of one or a few extremely high or low values has no effect upon the mode. If a series of data is unclassified, not having been either arrayed or put into a frequency distribution, the mode cannot be readily located.

For example, if seven men are receiving daily wages of Rs 5, 6, 7, 7, 7, 8 and 10, it is clear that the modal wage is Rs 7 per day. If you have a series, such as 2, 3, 5, 6, 7, 10 and 11, it is apparent that there is no mode.

There are several methods of estimating the value of the mode. But, it is seldom that the different methods of ascertaining the mode give us identical results. Consequently, it becomes necessary to decide as to which method would be most suitable for the purpose in hand. In order that a choice of the method may be

made, you should understand each of the methods and the differences that exist among them.

The four important methods of estimating mode of a series are: (*i*) Locating the most frequently repeated value in the array, (*ii*) estimating the mode by interpolation, (*iii*) locating the mode by graphic method and (*iv*) estimating the mode from the mean and the median. Only the last three methods are discussed in this unit.

Estimating the mode by interpolation

In the case of continuous frequency distributions, the problem of determining the value of the mode is not so simple as it might have appeared from the foregoing description. Having located the modal class of the data, the next problem in the case of continuous series is to interpolate the value of the mode within this 'modal' class.

The interpolation is made by the use of any one of the following formulae:

$$(i)\ Mo = l_1 + \frac{f_2}{f_0 + f_2} \times i;\quad (ii)\ Mo = l_2 - \frac{f_0}{f_0 + f_2} \times i$$

or $$(iii)\ Mo = l_1 + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times i$$

where l_1 is the lower limit of the modal class, l_2 is the upper limit of the modal class, f_0 equals the frequency of the preceding class in value, f_1 equals the frequency of the modal class in value, f_2 equals the frequency of the following class (class next to modal class) in value and i equals the interval of the modal class.

Example 2.13: Determine the mode for the data given in the following table.

Wage Group	*Frequency (f)*
14 – 18	6
18 – 22	18
22 – 26	19
26 – 30	12
30 – 34	5
34 – 38	4
38 – 42	3
42 – 46	2
46 – 50	1
50 – 54	0
54 – 58	1

Solution: In the given data, 22–26 is the modal class (since it has the largest frequency of 19), the lower limit of the modal class is 22, its upper limit is 26, the frequency of the preceding class is 18, and of the following class is 12. The class interval is 4. Using the various methods of determining mode, we have

NOTES

$$(i)\ Mo = 22 + \frac{12}{18+12} \times 4 \qquad (ii)\ Mo = 26 - \frac{18}{18+12} \times 4$$

$$= 22 + \frac{8}{5} \qquad = 26 - \frac{12}{5}$$

$$= 23.6 \qquad = 23.6$$

$$(iii)\ Mo = 22 + \frac{19-18}{(19-18)+(19-12)} \times 4 \qquad = 22 + \frac{4}{8} = 22.5$$

In formulae (*i*) and (*ii*), the frequencies of the classes adjoining the modal class are used to pull the estimate of the mode away from the midpoint towards either the upper or lower class limit. In this particular case, the frequency of the class preceding the modal class is more than the frequency of the class following and, therefore, the estimated mode is less than the mid-value of the modal class. This seems quite logical. If the frequencies are more on one side of the modal class than on the other, it can be reasonably concluded that the items in the modal class are concentrated more towards the class limit of the adjoining class with the larger frequency.

Formula (*iii*) is also based on a logic similar to that of (*i*) and (*ii*). In this case, to interpolate the value of the mode within the modal class, the differences between the frequency of the modal class and the respective frequencies of the classes adjoining it are used. This formula usually gives better results than the values obtained by the other two formulae. Also, the result given by this formula is the same as the one give by the graphic method. The formulae (*i*) and (*ii*) give values which are different from the value obtained by formula (*iii*) and are more close to the central point of modal class. If the frequencies of the classes adjoining the modal are equal, the mode is expected to be located at the mid-value of the modal class, but if the frequency on one of the sides is greater, the mode will be pulled away from the central point. It will be pulled more and more if the difference between the frequencies of the classes adjoining the modal class is higher and higher. In Example 2.13, the frequency of the modal class is 19 and that of the preceding class is 18. So, the mode should be quite close to the lower limit of the modal class. The midpoint of the modal class is 24 and the lower limit of the modal class is 22.

Locating the mode by the graphic method

The method of graphic interpolation is illustrated in Figure 2.4. The upper corners of the rectangle over the modal class are joined by straight lines to those of the adjoining rectangles as shown in the diagram; the right corner to the corresponding one of the adjoining rectangle on the left, etc. If a perpendicular is drawn from the points of intersection of these lines, you have a value for the mode indicated on the base line. The graphic approach is, in principle, similar to the arithmetic interpolation explained earlier.

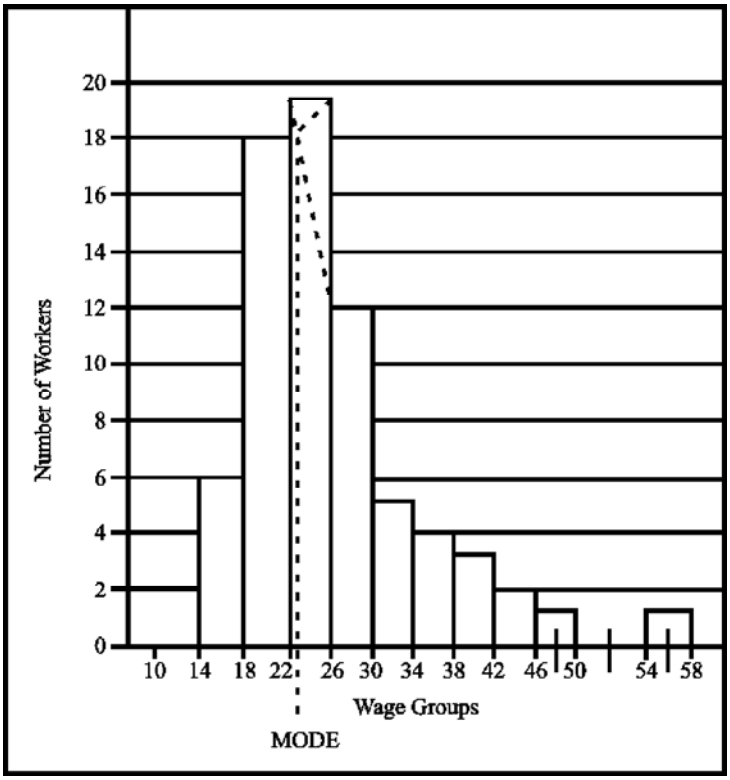


Fig. 2.4 *Method of Mode Determination by Graphic Interpolation*

The mode may also be determined graphically from an ogive or cumulative frequency curve. It is found by drawing a perpendicular to the base from that point on the curve where the curve is most nearly vertical, i.e. steepest (in other words, where it passes through the greatest distance vertically and the smallest distance horizontally). The point where it cuts the base will give you the value of the mode. How accurately this method determines the mode is governed by (*i*) the shape of the ogive and (*ii*) the scale on which the curve is drawn.

Estimating the mode from the mean and the median

There usually exists a relationship among the mean, median and mode for moderately asymmetrical distributions. If the distribution is symmetrical, the mean, median and mode will have identical values, but if the distribution is skewed (moderately), the mean, median and mode will pull apart. If the distribution tails off towards higher values, the mean and the median will be greater than the mode. If it tails off towards lower values, the mode will be greater than either of the other two measures. In either case, the median will be about one-third as far away from the mean as the mode is. This means that

$$\text{Mode} = \text{Mean} - 3\,(\text{Mean} - \text{Median})$$
$$= 3\,\text{Median} - 2\,\text{Mean}$$

In the case of the average monthly earnings (refer to the table of Example 2.3), the mean is 68.53 and the median is 70.2. If these values are substituted in the above formula, you get

$$\text{Mode} = 68.5 - 3(68.5 - 70.2)$$
$$= 68.5 + 5.1 = 73.6$$

According to the formula used earlier,

NOTES

$$\text{Mode} = l_1 + \frac{f_2}{f_0 + f_2} \times i$$
$$= 72.5 + \frac{745}{795 + 745} \times 5$$
$$= 72.5 + 2.4 = 74.9$$

OR

$$\text{Mode} = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$
$$= 72.5 + \frac{915 - 795}{2 \times 915 - 795 - 745} \times 5$$
$$= 72.5 + \frac{120}{290} \times 5 = 75.57$$

The difference between the two estimates is due to the fact that the assumption of relationship between the mean, median and mode may not always be true; it is obviously not valid in this case.

Example 2.14: (a) In a moderately symmetrical distribution, the mode and mean are 32.1 and 35.4 respectively. Calculate the median.

(b) If the mode and median of moderately asymmetrical series are respectively 16" and 15.7", what would be its most probable median?

(c) In a moderately skewed distribution, the mean and the median are respectively 25.6 and 26.1 inches. What is the mode of the distribution?

Solution: (a) You know:

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

or $$3\,\text{Median} = \text{Mode} + 2\,\text{Mean}$$

or $$\text{Median} = \frac{32.1 + 2 \times 35.4}{3}$$
$$= \frac{102.9}{3}$$
$$= 34.3$$

(b) $$2\,\text{Mean} = 3\,\text{Median} - \text{Mode}$$

or $$\text{Mean} = \frac{1}{2}(3 \times 15.7 - 16.0) = \frac{31.1}{2} = 15.55$$

(c) $$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$
$$= 3 \times 26.1 - 2 \times 25.6 = 78.3 - 51.2 = 27.1$$

CHECK YOUR PROGRESS

3. What is descriptive statistics?
4. What is the arithmetic mean.
5. What are the advantages of mean?
6. The following are the scores for the mid-term exam given to 13 students in Statistics.

   42, 42, 68, 80, 75, 54, 62, 89, 72, 80, 80, 75, 65

   Calculate:

   (a) The mean

   (b) The mode

   (c) The median
7. The following data represent the number of cars entering a gas station on Bedford Avenue for repairs between 10.00 AM and 11.00 AM in the last 8 days:

   7, 8, 6, 8, 9, 7, 5, 6

   Calculate the mean for the given data.
8. The following are the monthly salaries, in rupees, of the employees in a branch bank:

   10, 17, 29, 95, 95, 100, 100, 175, 250 and 750

   Calculate the arithmetic mean.
9. The following figures represent the number of books issued at the counter of a commerce library in 11 different days.

   96, 180, 98, 75, 270, 20, 102, 100, 94, 75, 200

   Calculate the median.

2.4 VARIATIONS

The square of standard deviation, namely σ^2, is termed as variance and is more often specified than the standard deviation. Clearly, it has the same properties as standard deviation.

As is clear, the standard deviation σ or its square, the variance, cannot be very useful in comparing two series where either the units are different or the mean values are different. Thus, a σ of 5 on an examination where the mean score is 30 has an altogether different meaning than on an examination where the mean score is 90. Clearly, the variability in the second examination is much less. To take care of this problem, we define and use a coefficient of variation, *V*,

$$V = \frac{\sigma}{\bar{x}} \times 100 \text{ expressed as percentage.}$$

NOTES

Example 2.15: The following are the scores of two batsmen *A* and *B* in a series of innings:

A	12	115	6	73	7	19	119	36	84	29
B	47	12	76	42	4	51	37	48	13	0

Who is the better run-getter? Who is more consistent?

Solution: In order to decide as to which of the two batsmen, *A* and *B*, is the better run-getter, you should find their batting averages. The one whose average is higher will be considered as a better batsman.

To determine the consistency in batting, you should determine the coefficient of variation. The less this coefficient the more consistent will be the player.

A			*B*		
Scores x	x	x^2	*Scores* x	x	x^2
12	–38	1,444	47	14	196
115	+65	4,225	12	–21	441
6	–44	1,936	76	43	1,849
73	+23	529	42	9	81
7	–43	1,849	–4	– 29	841
19	–31	961	51	18	324
119	+69	4,761	37	4	16
36	–14	196	48	15	225
84	+34	1,156	13	–20	400
29	–21	441	0	–33	1,089
$\sum x = 500$		17,498	$\sum x = 330$		5,462

Batsman *A*:

$$\bar{x} = \frac{500}{10} = 50$$

$$\sigma = \sqrt{\frac{17,498}{10}} = 41.83$$

$$V = \frac{41.83 \times 100}{50}$$

= 83.66 per cent

Batsman *B:*

$$\bar{x} = \frac{330}{10} = 33$$

$$\sigma = \sqrt{\frac{5,462}{10}} = 23.37$$

$$V = \frac{23.37}{33} \times 100$$

= 70.8 per cent

A is a better batsman since his average is 50 as compared to 33 of *B*. But *B* is more consistent since the variation in his case is 70.8 as compared to 83.66 of *A*.

Example 2.16: The following table gives the age distribution of students admitted to a college in the years 1914 and 1918. Find which of the two groups is more variable in age.

NOTES

Age	*Number of Students*	
	1914	*1918*
15	–	1
16	1	6
17	3	34
18	8	22
19	12	35
20	14	20
21	13	7
22	5	19
23	2	3
24	3	–
25	1	–
26	–	–
27	1	–

Solution:

Age	*Assumed Mean—2l 1914*				*Assumed Mean—19 1918*			
	f	*x'*	*fx'*	fx'^2	*f*	*x'*	*fx*	fx'^2
15	0	–6	0	0	1	–4	–4	16
16	1	–5	–5	25	6	–3	–18	54
17	3	–4	–12	48	34	–2	–68	136
18	8	–3	–24	72	22	–1	–22	22
19	12	–2	–24	48			–112	
20	14	–1	–14	14				
			–79		35	0	0	0
21	13	0	0	0	20	1	20	20
22	5	1	5	5	7	2	14	28
23	2	2	4	8	19	3	57	171
24	3	3	9	27	3	4	12	48
25	1	4	4	16	147		+103	495
26	0	5	0	0			–9	
27	1	6	6	36				
	63		+28	299				
			–51					

1914 Group:

$$\sigma = \sqrt{\frac{\sum fx'^2}{N} - \left[\frac{\sum (fx')}{N}\right]^2}$$

$$= \sqrt{\frac{299}{63} - \left(\frac{-51}{63}\right)^2}$$

NOTES

$$= \sqrt{4.476 - 0.655} = \sqrt{4.091}$$

$$= 2.02$$

$$\bar{x} = 21 + \left(\frac{-51}{63}\right) = 21 - 8 = 20.2$$

$$V = \frac{2.02}{20.2} \times 100$$

$$= \frac{202}{20.2} = 10$$

1918 Group:

$$\sigma = \sqrt{\frac{495}{147} - \left(\frac{-9}{147}\right)^2} = \sqrt{3.3673 - 0.0037}$$

$$= \sqrt{3.3636} = 1.834$$

$$\bar{x} = 19 + \left(\frac{-9}{147}\right)$$

$$= 19 - 0.06 = 18.94$$

$$V = \frac{1.834}{18.94} \times 100$$

$$= 9.68$$

The coefficient of variation of the 1914 group is 10 and that of the 1918 group 9.68. This means that the 1914 group is more variable, but only barely so.

Example 2.17: You are supplied the following data about the height of boys and girls studying in a college.

	Boys	Girls
Number	72	38
Average height (inches)	68	61
Variance of distribution	9	4

You are required to find out:

(a) In which sex, boys or girls, there is greater variability in individual heights.

(b) Common average height of boys and girls.

(c) Standard deviation of the height of boys and girls taken together.

(d) Combined variability (CV).

NOTES

Solution:

(a) CV of boys' height = $\frac{\sigma_1}{\bar{x}_1} \times 100 = \frac{\sqrt{9}}{68} \times 100 = 4.41\%$

CV of girls' height = $\frac{\sigma_2}{\bar{x}_2} \times 100 = \frac{\sqrt{4}}{61} \times 100 = 3.28\%$

Thus, there is a greater variability in the height of boys than that of the girls.

(b) The combined height of boys and girls is given as

$$\bar{x}_{12} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2}$$

$$= \frac{72 \times 68 + 38 \times 61}{72 + 38} = \frac{7214}{110} = 65.58 \text{ inches approx.}$$

(c) The combined standard deviation may be calculated by applying the following formula:

$$\sigma_{12}^2 = \frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2} + \frac{N_1(\bar{x} - \bar{x}_1)^2 + N_2(\bar{x} - \bar{x}_2)^2}{N_1 + N_2}$$

$$= \frac{72 \times 9 + 38 \times 4}{72 + 38} + \frac{72(65.58 - 68)^2 + 38(65.58 - 61)^2}{72 + 38}$$

$$= \frac{2018.794}{110} = 18.35$$

$$\sigma_{12} = 4.28 \text{ inches}$$

(d) Combined variability = $\frac{\sigma}{\bar{x}} \times 100 = \frac{4.28}{65.58} \times 100 = 6.53$

Example 2.18: In a co-educational college, boys and girls formed separate groups on the foundation day when everyone had to put in physical labour. Compute standard deviation for boys and girls separately and for the combined group. Did the separation by sex make each workgroup more homogeneous?

Minutes of Labour Given by Each Individual	*No. of Girls*	*No. of Boys*
60	20	120
55	60	100
50	100	200
45	450	355
40	450	350
35	300	500
30	250	350
25	100	20

Solution:

Minutes of Labour Given by Each Individual (X)	$X' = \frac{X-45}{5}$	*No. of Girls* f_1	f_1X'	$f_1X'^2$	*No. of Boys* f_2	f_2X'	$f_2X_2'^2$
60	+3	20	60	180	120	360	1080
55	+2	60	120	240	100	200	400
50	+1	100	100	100	200	200	200
45	0	450	0	0	355	0	0
40	–1	450	–450	450	350	–350	350
35	–2	300	–600	1200	500	–1000	2000
30	–3	250	–750	2250	350	–1050	3150
25	–4	100	–400	1600	20	–80	320
Total		1730	–1920	6020	1995	–1720	7500

Girls: $\bar{X}_1 = 45 - \frac{1920}{1730} \times 5 = 45 - 5.55 = 39.45$

$$\sigma_1 = \sqrt{\frac{6020}{1730} - \left(\frac{-1920}{1730}\right)^2} \times 5 = \sqrt{3.4798 - 1.2317} \times 5 = 1.5 \times 5 = 7.5$$

$$V_1 = \frac{7.5}{39.45} \times 100 = 19.0\%$$

Boys:

$$\bar{X}_2 = 45 - \frac{1720}{1995} \times 5 = 45 - 4.31 = 40.69$$

$$\sigma_2 = \sqrt{\frac{750}{1995} - \left(\frac{-1720}{1995}\right)^2} \times 5 = \sqrt{(3.7594 - 0.7434)} \times 5$$

$$= 1.7366 \times 5 = 8.68$$

$$V_2 = \frac{8.68}{40.69} \times 100$$

$$= 21.34\%$$

Combined mean: $\bar{X}_{12} = \frac{1730 \times 39.45 + 1995 \times 40.69}{1730 + 1995}$

$$= \frac{149424.78}{3725} = 40.11$$

$$d_1 = 0.66 \text{ and } d_2 = -0.58$$

$$\sigma_{12}^2 = \frac{1730\,(7.5)^2 + 1995\,(8.68)^2 + 1730\,(0.66)^2 + 1995\,(-0.58)^2}{1730 + 1995}$$

$$= 66.86$$

$$\sigma_{12} = 8.18$$

$$CV = \frac{8.18}{40.11} \times 100 = 20.38\%$$

Therefore, there does not seem to be any evidence that each workgroup was more homogeneous than the total population.

Example 2.19: The values of the arithmetic mean and the standard deviation of the following frequency distribution of a continuous variable derived from short-cut method are 135.3 lbs and 9.6 lbs respectively.

X	–4	–3	–2	–1	0	1	2	3	Total
Frequency	2	5	8	18	22	13	8	4	80

Determine the actual class interval.

Solution: Calculation of standard deviation:

X	–4	–3	–2	–1	0	1	2	3	Total
Frequency (*f*)	2	5	8	18	22	13	8	4	80
f(X)	–8	–15	–16	–18	0	13	16	12	–16
$f(X^2)$	32	45	32	18	0	13	32	36	208

$$\text{Standard Deviation} = i \times \sqrt{\frac{\sum f(X)}{n} - \left(\frac{\sum f(X)}{n}\right)^2}$$

$\therefore$ Putting the known values, we have

$$9.6 = i \times \sqrt{\frac{208}{80} - \left(\frac{-16}{80}\right)^2} \quad = i \times \sqrt{2.6 - 0.04}$$

or $$9.6 = i \times \sqrt{2.56} = i \times 1.6$$

$\therefore$ $$i = \frac{9.6}{1.6} = 6$$

$$\text{Arithmetic mean} = A + \frac{\Sigma f(X)}{n} \times i$$

$\therefore$ Putting the known values, we have

$$135.3 = A + \frac{-16}{80} \times 6 = A - 1.2$$

or $$A = 135.3 + 1.2 = 136.5$$

A, or assumed, mean is the midpoint corresponding to the class having *X* value 0. As the class interval is of 6 and the variable under studying is a continuous one, the class for which *X* = 0 will be 136.5 –3 to 136.5 + 3, i.e. 133.5–139.5. The class next lower than this is 133.5–6 or 133.5, i.e. 127.5 to 133.5.

Similarly, other classes can be calculated. So all the class intervals are:

109.5–115.5 115.5–121.5 121.5–127.5 127.5–133.5
133.5–139.5 139.5–145.5 145.5–151.5 151.5–157.5

Determining overall performance

If for a number of candidates each writes papers in several subjects, then the total marks obtained by candidates will not be a correct basis for determining their merit.

This is also because the marks in different subjects are likely to have different amounts of spread and means. The extent of the spread of marks in a paper introduces 'weights' which will be to the advantage of those getting high marks in subjects where the spread is great, and to the disadvantage of those gaining high marks in subjects where range is small.

A method to for the errors so introduced consists of converting the marks into 'standard scores' defined as $\frac{X - \bar{X}}{\sigma}$. This measures deviation of marks of a student from the mean of that subject in the units of standard deviation and are termed as *z*-scores as well. This corrects both for variations of $\bar{X}$ and σ in different subjects.

The *z*-scores of a student in different subjects are then added to give a true measure of relative performance.

Example 2.20: Consider the following data:

Candidate	*Marks in*		
	Economics	*Commerce*	*Total*
A	84	75	159
B	74	85	159

Average for Economics is 60 with standard deviation 13

Average for Commerce is 50 with standard deviation 11

Based on these information, determine whose performance is better *A*'s or *B*'s.

Solution:

z-Scores

A: Economics $\frac{84 - 60}{13} = 1.85$; Commerce $\frac{75 - 50}{11} = 2.27$ } 4.12

B: Economics $\frac{74 - 60}{13} = 1.08$; Commerce $\frac{85 - 50}{11} = 3.28$ } 4.26

Since *B*'s *z*-score is higher therefore his performance is better.

2.5 DISPERSION

2.5.1 Measures of Dispersion: Definition

A measure of dispersion, or simply *dispersion* may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency.

A measure of dispersion may be expressed in an 'absolute form', or in a 'relative form'. It is said to be in an absolute form when it states the actual amount by which the value of an item on an average deviates from a measure of central tendency. Absolute measures are expressed in concrete units, i.e. units in terms of which the data have been expressed, e.g. rupees, centimetres, kilograms, etc., and are used to describe frequency distribution.

A relative measure of dispersion is a quotient obtained by dividing the absolute measures by a quantity in respect to which absolute deviation has been computed. It is as such a pure number and is usually expressed in a percentage form. Relative measures are used for making comparisons between two or more distributions.

A measure of dispersion should possess the following characteristics which are considered essential for a measure of central tendency.

(a) It should be based on all observations.

(b) It should be readily comprehensible.

(c) It should be fairly and easily calculated.

(d) It should be affected as little as possible by fluctuations of sampling.

(e) It should be amenable to algebraic treatment.

The following are the common measures of dispersion:

(*i*)The range, (*ii*) The semi-interquartile range or the quartile deviation, (*iii*) The mean deviation and (*iv*) The standard deviation. Of these, the standard deviation is the best measure. All these measures are discussed in this unit.

2.5.2 The Range

The crudest measure of dispersion is the range of the distribution. The range of any series is the difference between the highest and the lowest values in the series. If the marks received in an examination taken by 248 students are arranged in ascending order, then the range will be equal to the difference between the highest and the lowest marks.

In a frequency distribution, the range is taken to be the difference between the lower limit of the class at the lower extreme of the distribution and the upper limit of the class at the upper extreme.

NOTES

Table 2.3 *Weekly Earnings of Labourers in Four Workshops of the Same Type*

Weekly earnings Rs	*No. of Workers*			
	Workshop A	*Workshop B*	*Workshop C*	*Workshop D*
15–16	...	...	2	...
17–18	...	2	4	...
19–20	...	4	4	4
21–22	10	10	10	14
23–24	22	14	16	16
25–26	20	18	14	16
27–28	14	16	12	12
29–30	14	10	6	12
31–32	...	6	6	4
33–34	...	...	2	2
35–36	...	...	...	...
37–38	...	...	4	...
Total	80	80	80	80
Mean	25.5	25.5	25.5	25.5

Consider the data on weekly earning of worker on four workshops given in Table 2.3. We note the following:

Workshop	*Range*
A	9
B	15
C	23
D	15

From these figures, it is clear that the greater the range, the greater is the variation of the values in the group.

The range is a measure of absolute dispersion and as such cannot be usefully employed for comparing the variability of two distributions expressed in different units. The amount of dispersion measured, say, in pounds, is not comparable with dispersion measured in inches. So the need of measuring relative dispersion arises.

An absolute measure can be converted into a relative measure if we divide it by some other value regarded as standard for the purpose. You may use the mean of the distribution or any other positional average as the standard.

From Table 2.3, the relative dispersion would be:

$$\text{Workshop } A = \frac{9}{25.5} \qquad \text{Workshop } C = \frac{23}{25.5}$$

$$\text{Workshop } B = \frac{15}{25.5} \qquad \text{Workshop } D = \frac{15}{25.5}$$

An alternate method of converting an absolute variation into a relative one would be to use the total of the extremes as the standard. This will be equal to dividing the difference of the extreme items by the total of the extreme items. Thus,

$$\text{Relative dispersion} = \frac{\text{Difference of extreme items, i.e., Range}}{\text{Sum of extreme items}}$$

The relative dispersion of the series is called the coefficient or ratio of dispersion. In our example of weekly earnings of workers considered earlier, the coefficients would be:

NOTES

$$\text{Workshop } A = \frac{9}{21+30} = \frac{9}{51} \qquad \text{Workshop } B = \frac{15}{17+32} = \frac{15}{49}$$

$$\text{Workshop } C = \frac{23}{15+38} = \frac{23}{53} \qquad \text{Workshop } D = \frac{15}{19+34} = \frac{15}{53}$$

Merits and limitations of range

Merits. Of the various characteristics that a good measure of dispersion should possess, the range has only two, namely (*i*) it is easy to understand, and (*ii*) its computation is simple.

Limitations. Besides the aforesaid two qualities, the range does not satisfy the other test of a good measure and hence it is often termed as a crude measure of dispersion.

The following are the limitations that are inherent in the range as a concept of variability:

(*i*) Since it is based upon two extreme cases in the entire distribution, the range may be considerably changed if either of the extreme cases happens to drop out, while the removal of any other case would not affect it at all.

(*ii*) It does not tell anything about the distribution of values in the series relative to a measure of central tendency.

(*iii*) It cannot be computed when distribution has open-end classes.

(*iv*) It does not take into account the entire data. These can be illustrated by the following illustration. Consider the data given in Table 2.4.

Table 2.4 *Distribution with the Same Number of Cases, but Different Variability*

Class	*No. of Students*		
	Section A	*Section B*	*Section C*
0–10	...	...	...
10–20	1	...	...
20–30	12	12	19
30–40	17	20	18
40–50	29	35	16
50–60	18	25	18
60–70	16	10	18
70–80	6	8	21
80–90	11	...	...
90–100	...	...	...
Total	110	110	110
Range	80	60	60

NOTES

The table is designed to illustrate three distributions with the same number of cases but different variability. The removal of two extreme students from section *A* would make its range equal to that of *B* or *C*.

The greater range of *A* is not a description of the entire group of 110 students, but of the two most extreme students only. Further, though sections *B* and *C* have the same range, the students in section B cluster more closely around the central tendency of the group than they do in section *C*. Thus, the range fails to reveal the greater homogeneity of *B* or the greater dispersion of *C*. Due to this defect, it is seldom used as a measure of dispersion.

Specific uses of range

In spite of the numerous limitations of the range as a measure of dispersion, there are the following circumstances when it is the most appropriate one:

- In situations where the extremes involve some hazard for which preparation should be made, it may be more important to know the most extreme cases to be encountered than to know anything else about the distribution. For example, an explorer, would like to know the lowest and the highest temperatures on record in the region he is about to enter; or an engineer would like to know the maximum rainfall during 24 hours for the construction of a storem water drain.
- In the study of prices of securities, range has a special field of activity. Thus to highlight fluctuations in the prices of shares or bullion it is a common practice to indicate the range over which the prices have moved during a certain period of time. This information, besides being of use to the operators, gives an indication of the stability of the bullion market, or that of the investment climate.
- In statistical quality control the range is used as a measure of variation. For example, you determine the range over which variations in quality are due to random causes, which is made the basis for the fixation of control limits.

2.5.3 Types of Measures

1. Quartile deviation (QD)

Another measure of dispersion, much better than the range, is the semi-interquartile range, usually termed as 'quartile deviation'. As you know, quartiles are the points which divide the array in four equal parts. More precisely, Q_1 gives the value of the item 1/4th the way up the distribution and Q_3 the value of the item 3/4th the way up the distribution. Between Q_1 and Q_3 are included half the total number of items. The difference between Q_1 and Q_3 includes only the central items but excludes the extremes. Since under most circumstances, the central half of the series tends to be fairly typical of all the items, the interquartile range ($Q_3 - Q_1$) affords a convenient and often a good indicator of the absolute variability. The larger the interquartile range, the larger the variability.

Usually, one-half of the difference between Q_3 and Q_1 is used and to it is given the name of quartile deviation or semi-interquartile range. The interquartile range is divided by two for the reason that half of the interquartile range will, in a normal distribution, be equal to the difference between the median and any quartile. This means that 50 per cent items of a normal distribution will lie within the interval defined by the median plus and minus the semi-interquartile range.

Symbolically,

$$QD = \frac{Q_3 - Q_1}{2}$$

Let us find quartile deviations for the weekly earnings of labour in the four workshop whose data is given in Table 2.3. The computations are given in Table 2.5.

As shown in the table, QD of workshop *A* is Rs 2.12 and median value is 25.3. This means that if the distribution is symmetrical, the number of workers whose wages vary between (25.3–2.1) = Rs 23.2 and (25.3 + 2.1) = Rs 27.4, shall be just half of the total cases. The other half of the workers will be more than Rs 2.1 removed from the median wage. As this distribution is not symmetrical, the distance between Q_1 and the median Q_2 is not the same as between Q_3 and the median. Hence, the interval defined by median plus and minus semi inter-quartile range will not be exactly the same as given by the value of the two quartiles. Under such conditions, the range between Rs 23.2 and Rs 27.4 will not include precisely 50 per cent of the workers.

If QD is to be used for comparing the variability of any two series, it is necessary to convert the absolute measure to a coefficient of quartile deviation. To do this the absolute measure is divided by the average size of the two quartile.

Symbolically,

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Applying this to your illustration of four workshops, the coefficients of QD would be as given in Table 2.5.

Table 2.5 *Calculation of Quartile Deviation*

	Workshop A	*Workshop B*	*Workshop C*	*Workshop D*
Location of Q_2 $\frac{N}{2}$	$\frac{80}{2} = 40$	$\frac{80}{2} = 40$	$\frac{80}{2} = 40$	$\frac{80}{2} = 40$
Q_2	$24.5 + \frac{40-30}{22} \times 2$ = 24.5 + 0.9 = 25.4	$24.5 + \frac{40-30}{18} \times 2$ = 24.5 + 1.1 = 25.61	$24.5 + \frac{40-30}{16} \times 2$ = 24.5 + 0.75 = 25.25	$24.5 + \frac{40-30}{16} \times 2$ = 24.5 + 0.75 = 25.25

Location of Q_1 $\frac{N}{4}$	$\frac{80}{4}=20$	$\frac{80}{4}=20$	$\frac{80}{4}=20$	$\frac{80}{4}=20$
Q_1	$22.5+\frac{20-10}{22}\times 2$ = 22.5 + 0.91 = 23.41	$22.5+\frac{20-16}{14}\times 2$ = 22.5 + 0.57 = 23.07	$20.5+\frac{20-10}{10}\times 2$ = 20.5 + 2 = 22.5	$22.5+\frac{20-18}{16}\times 2$ = 22.5 + 0.25 = 22.75
Location of Q_3 $\frac{3N}{4}$	$3\times\frac{80}{4}=60$	60	60	60
Q_3	$26.5+\frac{60-52}{14}\times 2$ = 26.5 + 1.14 = 27.64	$26.5+\frac{60-48}{16}\times 2$ = 26.5 + 1.5 = 28.0	$26.5+\frac{60-50}{12}\times 2$ = 26.5 + 1.67 = 28.17	$26.5+\frac{60-50}{12}\times 2$ = 26.5 + 1.67 = 28.17
Quartile Deviation $\frac{Q_3-Q_1}{2}$	$\frac{27.64-23.41}{2}$ $=\frac{4.23}{2}$ = Rs 2.12	$\frac{28-23.07}{2}$ $=\frac{4.93}{2}$ = Rs 2.46	$\frac{28.17-22.5}{2}$ $=\frac{5.67}{2}$ = Rs 2.83	$\frac{28.17-22.75}{2}$ $=\frac{5.42}{2}$ = Rs 2.71
Coefficient of Quartile Deviation $\frac{Q_3-Q_1}{Q_3+Q_1}$	$=\frac{27.64-23.41}{27.64+23.41}$ = 0.083	$\frac{28-23.07}{28+23.07}$ = 0.097	$\frac{28.17-22.5}{28.17+22.5}$ = 0.112	$\frac{28.17-22.75}{28.17+22.75}$ = 0.106

Characteristics of QD

The characteristics of quartile deviation are as follows:

(*i*) The size of QD gives an indication about the uniformity or otherwise of the size of the items of a distribution. If QD is small, it denotes large uniformity. Thus, a coefficient of QD may be used for comparing uniformity or variation in different distributions.

(*ii*) QD is not a measure of dispersion in the sense that it does not show the scatter around an average, but it shows only a distance on scale. Consequently, quartile deviation is regarded as a measure of partition.

(*iii*) QD can be computed when the distribution has open-end classes.

Limitations of QD

Except for the fact that its computation is simple and it is easy to understand, a QD does not satisfy any other test of a good measure of variation.

2. Mean deviation (MD)

A weakness of the measures of dispersion, based upon the range or a portion thereof, is that the precise size of most of the variants has no effect on the result. As an illustration, QD will be the same whether the variates between Q_1 and Q_3 are concentrated just above Q_1 or they are spread uniformly from Q_1 to Q_3. This is an important defect from the viewpoint of measuring the divergence of the distribution from its typical value. The mean deviation is employed to overcome this problem.

Mean deviation, also called average deviation, of a frequency distribution is the mean of the absolute values of the deviation from some measure of central tendency. In other words, mean deviation is the arithmetic average of the variations (deviations) of the individual items of the series from a measure of their central tendency.

You can measure the deviations from any measure of central tendency, but the most commonly employed ones are the median and the mean. The median is preferred because it has the important property that the average deviation from it is the least.

Calculation of the mean deviation then involves the following steps:

(a) Calculate the median or the mean, *Md* or *Me* ($\bar{x}$).

(b) Record the deviations $| d | = | x - Me |$ of each of the items, ignoring the sign.

(c) Find the average value of deviations.

$$\text{Mean deviation} = \frac{\Sigma | d |}{N}$$

Example 2.21: Calculate the mean deviation from the following data giving marks obtained by 11 students in a class test.

14, 15, 23, 20, 10, 30, 19, 18, 16, 25, 12

Solution: Median = Size of $\frac{11+1}{2}$ th item

= Size of 6th item = 18

Serial No.	*Marks*	$\| x - Median \|$ $\| d \|$
1	10	8
2	12	6
3	14	4
4	15	3
5	16	2
6	18	0
7	19	1
8	20	2
9	23	5
10	25	7
11	30	12
		$\Sigma \| d \| = 50$

$$\text{Mean deviation from median} = \frac{\Sigma | d |}{N}$$

$$= \frac{50}{11} = 4.5 \text{ marks}$$

NOTES

For grouped data, it is easy to see that the mean deviation is given by

$$\text{Mean Deviation (MD)} = \frac{\sum f\,|d|}{\sum f}$$

where $|d| = |x - \text{median}|$ for grouped discrete data, and $|d| = M - \text{median}|$ for grouped continuous data with M as the mid-value of a particular group. The following examples illustrate the use of this formula.

Example 2.22: Calculate the mean deviation from the following data:

Size of Item	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

Solution:

Size	*Frequency* (f)	*Cumulative Frequency*	*Deviations from Median* $\lvert d \rvert$	$f \lvert d \rvert$
6	3	3	3	9
7	6	9	2	12
8	9	18	1	9
9	13	31	0	0
10	8	39	1	8
11	5	44	2	10
12	4	48	3	12
	48			60

$$\text{Median} = \text{Size of } \frac{48+1}{2} = 24.5\text{th item which is } 9$$

Therefore, deviations (d) are calculated from 9, i.e. $|d| = |x - 9|$

$$\text{Mean deviation} = \frac{\sum f\,|d|}{\sum f} = \frac{60}{48} = 1.25$$

Example 2.23: Calculate the mean deviation from the following data:

x	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
f	18	16	15	12	10	5	2	2

NOTES

Solution: This is a frequency distribution with continuous variable. Thus, deviations are calculated from mid-values.

x	*Mid-value*	*f*	*Less than cf*	*Deviation from Median | d |*	*f | d |*
0–10	5	18	18	19	342
10–20	15	16	34	9	144
20–30	25	15	49	1	15
30–40	35	12	61	11	132
40–50	45	10	71	21	210
50–60	55	5	76	31	155
60–70	65	2	78	41	82
70–80	75	2	80	51	102
		80			1182

$$\text{Median} = \text{Size of } \frac{80}{2} \text{ th item}$$

$$= 20 + \frac{6}{15} \times 10 = 24$$

and then

$$\text{Mean deviation} = \frac{\Sigma f |d|}{\Sigma f}$$

$$= \frac{1182}{80} = 14.775$$

Merits and demerits of the mean deviation

Merits: The merits of the mean deviation are as follows:

(*i*) It is easy to understand.

(*ii*) As compared to standard deviation (discussed later), its computation is simple.

(*iii*) As compared to standard deviation, it is less affected by extreme values.

(*iv*) Since it is based on all values in the distribution, it is better than range or quartile deviation.

Demerits: The demerits of the mean deviation are as follows:

(*i*) It lacks those algebraic properties which would facilitate its computation and establish its relation to other measures.

(*ii*) Due to the above factor, it is not suitable for further mathematical processing.

Coefficient of mean deviation

The coefficient or relative dispersion is found by dividing the mean deviations (if deviations are recorded either from the mean or from the median) by mean or by median. Thus,

NOTES

$$\text{Coefficient of mean deviation} = \frac{\text{Mean Deviation}}{\text{Mean}}$$

(when deviations are recorded from the mean)

$$= \frac{\text{Mean Deviation}}{\text{Median}}$$

(when deviations are recorded from the median)

Applying the preceding formula to Example 2.23,

$$\text{Coefficient of mean deviation} = \frac{14.775}{24} = 0.616$$

3. Standard deviation

By far the most universally used and the most useful measure of dispersion is the standard deviation (SD) or root mean square deviation about the mean. You have seen that all the methods of measuring dispersion so far discussed are not universally adopted for want of adequacy and accuracy. The range is not satisfactory as its magnitude is determined by most extreme cases in the entire group. Further, the range is notable because it is dependent on the item whose size is largely matter of chance. The mean deviation method is also an unsatisfactory measure of scatter, as it ignores the algebraic signs of deviation. We desire a measure of scatter which is free from these shortcomings. To some extent, SD is one such measure.

The calculation of SD differs in the following respects from that of mean deviation. First, in calculating SD, the deviations are squared. This is done so as to get rid of negative signs without committing algebraic violence. Further, the squaring of deviations provides added weight to the extreme items, a desirable feature for certain types of series.

Secondly, the deviations are always recorded from the arithmetic mean, because although the sum of deviations is the minimum from the median, the sum of squares of deviations is minimum when deviations are measured from the arithmetic average. The deviation from $\bar{x}$ is represented by d.

Thus, SD, σ (sigma) is defined as the square root of the mean of the squares of the deviations of individual items from their arithmetic mean.

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{N}} \tag{2.5}$$

For grouped data (discrete variables),

$$\sigma = \sqrt{\frac{\sum f (x - \bar{x})^2}{\sum f}} \tag{2.6}$$

and, for grouped data (continuous variables),

$$\sigma = \sqrt{\frac{\sum f (M - \bar{x})}{\sum f}} \tag{2.7}$$

where M is the mid-value of the group.

The use of these equations is illustrated by the following examples.

Example 2.24: Compute the standard deviation for the following data:

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21

Solution: Here equation (2.5) is appropriate. We first calculate the mean as $\bar{x} = \sum x/N = 176/11 = 16$, and then calculate the deviation as follows:

x	$(x - \bar{x})$	$(x - \bar{x})^2$
11	–5	25
12	–4	16
13	–3	9
14	–2	4
15	–1	1
16	0	0
17	+1	1
18	+2	4
19	+3	9
20	+4	16
21	+5	25
176		11

Thus by formula (2.5)

$$\sigma = \sqrt{\frac{110}{11}} = \sqrt{10} = 3.16$$

Example 2.25: Find the standard deviation of the data in the following distributions:

x	12	13	14	15	16	17	18	20
f	4	11	32	21	15	8	6	4

Solution: For this discrete variable grouped data, you use equation (2.6). Since for calculation of $\bar{x}$, you need $\sum fx$ and then for σ you need $\sum f(x - \bar{x})^2$, the calculations are conveniently made in the following format.

x	f	fx	$d = x - \bar{x}$	d^2	fd^2
12	4	48	–3	9	36
13	11	143	–2	4	44
14	32	448	–1	1	32
15	21	315	0	0	0
16	15	240	1	1	15
17	8	136	2	4	32
18	5	90	3	9	45
20	4	80	5	25	100
	100	1500			304

Here $\bar{x} = \sum fx/\sum f = 1500/100 = 15$

and $\sigma = \sqrt{\frac{\sum fd^2}{\sum f}} = \sqrt{\frac{304}{100}} = \sqrt{3.04} = 1.74$

NOTES

Example 2.26: Calculate the standard deviation of the following data:

Class	1–3	3–5	5–7	7–19	9–11	11–13	13–15
Frequency	1	9	25	35	17	10	3

Solution: This is an example of continuous frequency series and equation (2.7) seems appropriate.

Class	*Mid-point* (x)	*Frequency* (f)	$f(x)$	*Deviation of Mid-point* (x) *from Mean* (d)	*Squared Deviation* d^2	*Squared Deviation Times Frequency* fd^2
1–3	2	1	2	–6	36	36
3–5	4	9	36	–4	16	144
5–7	6	25	150	–2	4	100
7–9	8	35	280	0	0	0
9–11	10	17	170	2	4	68
11–13	12	10	120	4	16	160
13–15	14	3	42	6	36	108
		100	800			616

First the mean is calculated as:

$$\bar{x} = \Sigma fx/\Sigma f = 800/100 = 8.0$$

Then the deviations are obtained from 8.0.

Thus, $$\sigma = \sqrt{\frac{\sum f(M - \bar{x})^2}{\sum f}}$$

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f}} = \sqrt{\frac{616}{100}} = 2.48$$

Calculation of SD by short-cut method

The three examples worked out above have one common simplifying feature, namely $\bar{x}$ in each case turned out to be an integer, thereby, simplifying calculations. In most cases, it is very unlikely that it will turn out to be so. In such cases, the calculation of d and d^2 becomes quite time-consuming. Thus, short-cut methods have been developed. These are on the same lines as those for calculation of mean itself.

In the short-cut method, you have to calculate deviations x' from an assumed mean A. Then, for ungrouped data,

$$\sigma = \sqrt{\frac{\sum x'^2}{N} - \left(\frac{\sum x'}{N}\right)^2} \quad (2.8)$$

and for grouped data,

$$\sigma = \sqrt{\frac{\sum f x'^2}{\sum f} - \left(\frac{\sum f x'}{\sum f}\right)^2} \quad (2.9)$$

NOTES

These formulae are valid for both discrete and continuous variables. In case of continuous variables, x in the equation $x' = x - A$ stands for the mid-value of the class in question.

Note that the second term in each of the formulae is a correction term because of the difference in the values of A and $\bar{x}$. When A is taken as $\bar{x}$ itself, this correction is automatically reduced to zero. The following examples explain the use of these formulae.

Example 2.27: Compute the standard deviation by the short-cut method for the following data:

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21

Solution: Let us assume that $A = 15$

	$x' = (x - 15)$	x'^2
11	–4	16
12	–3	9
13	–2	4
14	–1	1
15	0	0
16	1	1
17	2	4
18	3	9
19	4	16
20	5	25
21	6	36
$N = 11$	$\sum x' = 11$	$\sum x'^2 = 121$

$$\sigma = \sqrt{\frac{\sum x'^2}{N} - \left(\frac{\sum x'}{N}\right)^2}$$

$$= \sqrt{\frac{121}{11} - \left(\frac{11}{11}\right)^2} \qquad = \sqrt{11 - 1}$$

$$= \sqrt{10} \qquad = 3.16$$

Another method: If you assume A as zero, then the deviation of each item from the assumed mean is the same as the value of the item itself. Thus, 11 deviates from the assumed mean of zero by 11, 12 deviates by 12, and so on. As such, you work with deviations without having to compute them, and the formula takes the following shape:

x	x^2
11	121
12	144
13	169
14	196
15	225
16	256
17	289
18	324
19	361
20	400
21	441
176	2926

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$$

$$= \sqrt{\frac{2926}{11} - \left(\frac{176}{11}\right)^2} = \sqrt{266 - 256} = 3.16$$

Example 2.28: Calculate the standard deviation of the following data by the short-cut method.

Person	1	2	3	4	5	6	7
Monthly Income (Rupees)	300	400	420	440	460	480	580

Solution: In this data set, the values of the variables are very large making calculations cumbersome. It is advantageous to take a common factor out. Thus, you need to use $x' = \frac{x - A}{20}$. The standard deviation is calculated using x' and then the true value of σ is obtained by multiplying back by 20. The effective formula used is:

$$\sigma = C \times \sqrt{\frac{\sum x'^2}{N} - \left(\frac{\sum x'}{N}\right)^2}$$

Where C represents the common factor.

Using $x' = (x - 420)/20$

x	*Deviation from Assumed Mean* $x' = (x - 420)$	x'	x'^2
300	–120	–6	36
400	–20	–1	1
420	0	0	0
		–7	
440	20	1	1
460	40	2	4

480	60	3	9
580	160	8	64
		+14	
$N = 7$		$\Sigma x' = 7$	$\Sigma x'^2 = 115$

$$\sigma = 20 \times \sqrt{\frac{\sum x'^2}{N} - \left(\frac{\sum x'}{N}\right)^2} = 20\sqrt{\frac{115}{7} - \left(\frac{7}{7}\right)^2} = 78.56$$

Example 2.29: Calculate the standard deviation from the following data:

Size	6	9	12	15	18
Frequency	7	12	19	10	2

Solution:

x	*Frequency* (*f*)	*Deviation from Assumed Mean 12*	*Deviation divided by Common Factor 3* (*x'*)	*x' times Frequency* (*fx'*)	*x'² times frequency* (fx'^2)
6	7	–6	–2	–14	28
9	12	–3	–1	–12	12
12	19	0	0	0	0
15	10	3	1	10	10
18	2	6	2	4	8
	$N = 50$			$\sum fx' = -12$	$\sum fx'^2 = 58$

Since deviations have been divided by a common factor, you use

$$\sigma = C\sqrt{\frac{\sum fx'^2}{N} - \left(\frac{\sum fx'}{N}\right)^2}$$

$$= 3\sqrt{\frac{58}{50} - \left(\frac{-12}{50}\right)^2}$$

$$= 3\sqrt{1.1600 - 0.0576}$$

$$= 3 \times 1.05 = 3.15$$

Example 2.30: Obtain the mean and standard deviation of the first *N* natural numbers, i.e. of 1, 2, 3, ..., $N-1$, *N*.

Solution: Let *x* denote the variable which assumes the values of the first *N* natural numbers.

Then,

$$\bar{x} = \frac{\sum_{1}^{N} x}{N} = \frac{\frac{N(N+1)}{2}}{N} = \frac{N+1}{2}$$

NOTES

NOTES

Hence, $\sum_{1}^{N} x = 1 + 2 + 3 + \cdots + (N-1) + N$

$$= \frac{N(N+1)}{2}$$

To calculate the standard deviation σ, you use 0 as the assumed mean *A*. Then,

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$$

But, $\sum x^2 = 1^2 + 2^2 + 3^2 + \cdots + (N-1)^2 + N^2 = \frac{N(N+1)(2N+1)}{6}$

Therefore,

$$\sigma = \sqrt{\frac{N(N+1)(2N+1)}{6N} - \frac{N^2(N+1)^2}{4N^2}}$$

$$= \sqrt{\frac{(N+1)}{2}\left[\frac{2N+1}{3} - \frac{N+1}{2}\right]} = \sqrt{\frac{(N+1)(N-1)}{12}}$$

Thus for first 11 natural numbers,

$$\bar{x} = \frac{11+1}{2} = 6$$

and $\sigma = \sqrt{\frac{(11+1)(11-1)}{12}} = \sqrt{10} = 3.16$

Example 2.31:

	Mid-point (*x*)	*Frequency* (*f*)	*Deviation from Class of Assumed Mean* (*x′*)	*Deviation time Frequency* (*fx′*)	*Squared Deviation times Frequency* (fx'^2)
0–10	5	18	–2	–36	72
10–20	15	16	–1	–16	16
				–52	
20–30	25	15	0	0	0
30–40	35	12	1	12	12
40–50	45	10	2	20	40
50–60	55	5	3	15	45
60–70	65	2	4	8	32
70–80	75	1	5	5	25
				60	
		$\Sigma f = 79$		60	242
				–52	
				$\sum fx' = 8$	

NOTES

Solution: Since the deviations are from assumed mean and expressed in terms of class interval units,

$$\sigma = i \times \sqrt{\frac{\sum x'^2}{N} - \left(\frac{\sum fx'}{N}\right)^2}$$

$$= 10 \times \sqrt{\frac{242}{79} - \left(\frac{8}{79}\right)^2}$$

$$= 10 \times 1.75 = 17.5$$

Combining standard deviations of two distributions

If you are given two sets of data of N_1 and N_2 items with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations σ_1 and σ_2 respectively, you can obtain the mean and standard deviation $\bar{x}$ and σ of the combined distribution using the following formulae:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} \tag{2.10}$$

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\bar{x} - \bar{x}_1)^2 + N_2(\bar{x} - \bar{x}_2)^2}{N_1 + N_2}} \tag{2.11}$$

Example 2.32: The mean and standard deviations of two distributions of 100 and 150 items are 50, 5 and 40, 6 respectively. Find the standard deviation of all taken together.

Solution: Combined mean,

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} = \frac{100 \times 50 + 150 \times 40}{100 + 150} = 44$$

Combined standard deviation,

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\bar{x} - \bar{x}_1)^2 + N_2(\bar{x} - \bar{x}_2)^2}{N_1 + N_2}}$$

$$= \sqrt{\frac{100 \times (5)^2 + 150\,(6)^2 + 100\,(44 - 50)^2 + 150\,(44 - 40)^2}{100 + 150}}$$

$$= 7.46$$

Example 2.33: A distribution consists of three components with 200, 250, 300 items having mean 25, 10 and 15 and standard deviation 3, 4 and 5, respectively. Find the standard deviation of the combined distribution.

Solution: In the usual notations, here you have

$N_1 = 200,\ N_2 = 250,\ N_3 = 300$

$\bar{x}_1 = 25,\ \bar{x}_2 = 10,\ \bar{x}_3 = 15$

NOTES

Formulae (2.6) and (3.7) can easily be extended for combination of three series as:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{N_1 + N_2 + N_3}$$

$$= \frac{200 \times 25 + 250 \times 10 + 300 \times 15}{200 + 250 + 300}$$

$$= \frac{12000}{750} = 16$$

and,

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1(\bar{x} - \bar{x}_1)^2 + N_2(\bar{x} - \bar{x}_2)^2 + N_3(\bar{x} - \bar{x}_3)^2}{N_1 + N_2 + N_3}}$$

$$= \sqrt{\frac{200 \times 9 + 250 \times 16 + 300 \times 25 + 200 \times 81 + 250 \times 36 + 300 \times 1}{200 + 250 + 300}}$$

$$= \sqrt{51.73} = 7.19$$

2.5.4 Merits, Limitations and Characteristics of Measures

The range is the easiest to calculate the measure of dispersion, but since it depends on extreme values, it is extremely sensitive to the size of the sample, and to the sample variability. In fact, as the sample size increases, the range increases dramatically, because the more the items one considers, the more likely it is that some item will turn up which is larger than the previous maximum or smaller than the previous minimum. So, it is, in general, impossible to interpret properly the significance of a given range unless the sample size is constant. It is for this reason that there appears to be only one valid application of the range, namely in statistical quality control where the same sample size is repeatedly used, so that comparison of ranges is not distorted by differences in sample size.

The quartile deviations and other such positional measures of dispersions are also easy to calculate. However, they suffer from the disadvantage that they are not amenable to algebraic treatment. Similarly, the mean deviation is not suitable because we cannot obtain the mean deviation of a combined series from the deviations of component series. However, it is easy to interpret and easier to calculate than the standard deviation.

The standard deviation of a set of data is one of the most important statistics describing it. It lends itself to rigorous algebraic treatment, is rigidly defined and is based on all observations. It is, therefore, quite insensitive to sample size (provided the size is 'large enough') and is least affected by sampling variations.

It is used extensively in testing of hypothesis about population parameters based on sampling statistics.

In fact, the standard deviation has such stable mathematical properties that it is used as a standard scale for measuring deviations from the mean. If you are told that the performance of an individual is 10 points better than the mean, it really does not tell you enough, for 10 points may or may not be a large enough difference to be of significance. But if you know that σ for the score is only 4 points, so that on this scale, the performance is 2.5 σ better than the mean, the statement becomes meaningful. This indicates an extremely good performance. This sigma scale is a very commonly used scale for measuring and specifying deviations which immediately suggest the significance of the deviation.

The only disadvantage of the standard deviation lie in the amount of work involved in its calculation and the large weight it attaches to extreme values because of the process of squaring involved in its calculations.

CHECK YOUR PROGRESS

10. Coefficients of variation of two series are 58% and 69%. Their standard deviations are 21.2 and 15.6. What are their arithmetic means?
11. What is the absolute measure of dispersion?
12. What is the relative measure of dispersion?
13. Define range.

2.6 SKEWNESS

When a frequency distribution is not symmetrical, it is said to be asymmetrical or skewed. The nature of symmetry and the various types of asymmetry are illustrated in the given example.

Table 2.6 lists the heights of the students of a college.

Table 2.6 *Heights of the Students*

Class Interval	*A* *f*	*B* *f*	*C* *f*	*D* *f*
56.5–58.5	5	3	0	4
58.5–60.5	25	5	4	8
60.5–62.5	15	20	40	20
62.5–64.5	10	44	24	24
64.5–66.5	15	20	20	40
66.5–68.5	25	5	8	4
68.5–70.5	5	3	4	0
N	100	100	100	100
Mean (Me)	63.5	63.5	63.5	63.5
Median (Md)	63.5	63.5	63	64
Mode (Mo)	—	63.5	61.9	65.1

NOTES

The histograms and the corresponding curves are shown in Figures 2.5 and 2.6.

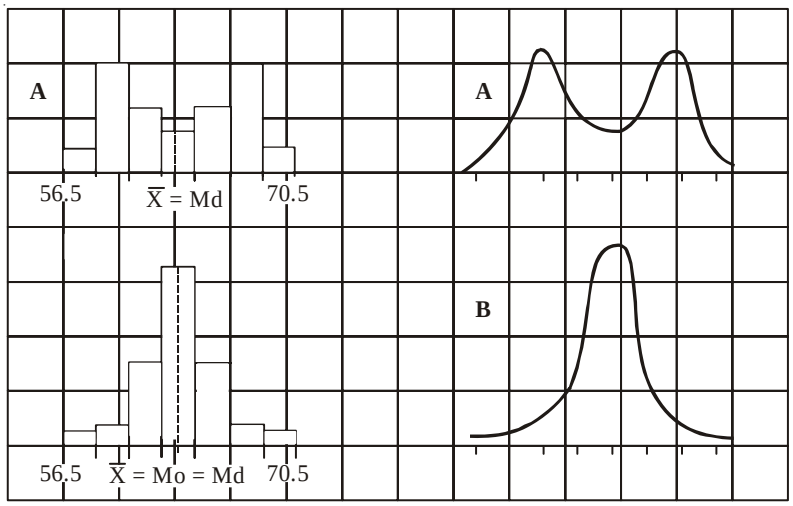


Fig. 2.5 *Histogram*

A glance at the data of each of the four classes, A, B, C and D makes a very interesting study.

The shape of the curves, histograms and placement of equal items at equal distances on either side of the median clearly show that distributions *A* and *B* are symmetrical. If you fold these curves, or histograms on the ordinate at the mean, the two halves of the curve or histograms will coincide. In distribution *B*, all the three measures of central tendency are identical. In *A*, which is a bimodal distribution, mean and median have the same value.

Distributions *C* and *D* are asymmetrical. This is evident from the shape of the histograms and curves, and also from the fact that items at equal distances from the median are not equal in number. The three measures of central tendency for each of these distributions are of different sizes.

A point of difference between the asymmetry of distribution *C* and that of *D* should be carefully noted. In distribution *C*, where the mean (63.5) is greater than the median (63) and the mode (61.9), the curve is pulled more to the right. In distribution *D*, where mean (63.5) is lesser than the median (64) and mode (65.1), the curve is pulled more to the left.

In other words, you may say that if the extreme variations in a given distribution are towards higher values, they give the curve a longer tail to the right and this pulls the median and mean in that direction from the mode. If, however, extreme variations are towards lower values, the longer tail is to the left and the median and mean are pulled to the left of the mode.

It could also be shown that in a symmetrical distribution, the lower and upper quartiles are equidistant from the median, so also are corresponding pairs of deciles and percentiles. This means that in a asymmetrical distribution, the distance of the upper and lower quartiles from median is unequal.

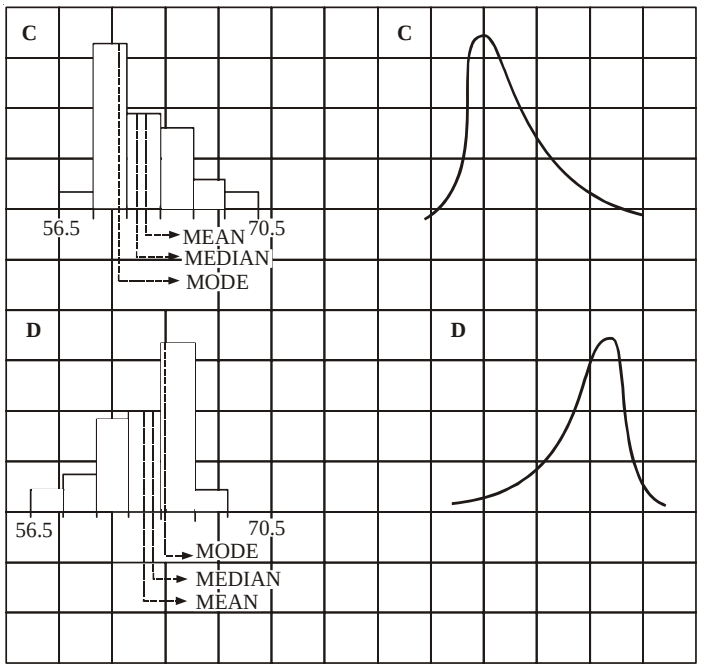


Fig. 2.6 *Skewness*

From the preceding discussion, we can summarize that skewness is present:

1. When the graph of the distribution does not show a symmetrical curve.
2. When the three measures of central tendency differ from one another.
3. When the sum of the positive deviations from the median is not equal to the negative deviations from the same value.
4. When the distances from the median to the quartiles are unequal.
5. When corresponding pairs of deciles or percentiles are not equidistant from the median.

2.6.1 Measures of Skewness

On the basis of the above tests, the following measures of skewness have been developed:

1. Relationship between three measures of central tendency—commonly known as the Karl Pearson's measure of skewness.
2. Quartile measure of skewness—known as Bowley's measure of skewness.
3. Percentile measure of skewness—also called the Kelly's measure of skewness.
4. Measures of skewness based on moments.

All these measures describe both the direction and the extent of the skewness.

1. Karl Pearson's measure of skewness

You know that in a perfectly symmetrical distribution, the three measures of central tendency, namely mean, median and mode will coincide. As the distribution departs from symmetry, these three values are pulled apart, the difference between the mean and mode being the greatest. Karl Pearson has suggested the use of this

NOTES

difference in measuring skewness. Thus, Absolute Skewness = Mean – Mode. (+) or (–) signs obtained by this formula would exhibit the direction of the skewness. If it is positive, the extreme variation in the given distribution is towards higher values. If it is negative, it shows that extreme variations are towards lower values.

Pearsonian coefficient of skewness

The difference between mean and mode, as explained in the preceding paragraph, is an absolute measure of skewness. An absolute measure cannot be used for making valid comparison between the skewness in two or more distributions for the following reasons: (*i*) the same size of skewness has different significance in distributions with small variation and in distributions with large variation, in the two series and (*ii*) the unit of measurement in the two series may be different.

To make this measure a suitable device for comparing skewness, it is necessary to eliminate from it the disturbing influence of 'variation' and 'units of measurements'. Such elimination is accomplished by dividing the difference between mean and mode by the standard deviation. The resultant coefficient is called *Pearsonian Coefficient of Skewness*. Thus, the formula of Pearsonian Coefficient of Skewness is:

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

You know that in moderately skewed distributions:

$$\text{Mode} = \text{Mean} - 3\,(\text{Mean} - \text{Median})$$

You may remove the mode from the formula by substituting the above in the formula for skewness as follows:

$$\text{Coefficient of skewness} = \frac{\text{Mean} - [\text{Mean} - 3(\text{Mean} - \text{Median})]}{\text{Standard Deviation}}$$

$$= \frac{\text{Mean} - \text{Mean} + 3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$= \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

The removal of the mode and substituting median in its place becomes necessary because mode cannot always be easily located and it is so much affected by grouping errors that it becomes unreliable.

Example 2.34: Find the skewness from the following data:

Height (*in inches*)	58	59	60	61	62	63	64	65
Number of Persons	10	18	30	42	35	28	16	8

Solution: Height is a continuous variable, and hence 58″ must be treated as 57.5″–58.5″, 59″ as 58.5″–59.5″, and so on.

Height (*in inches*)	*Frequency* *f*	*x′* *from 61*	*fx′*	*fx′*^{2}	*Cumulative Frequency*
58	10	–3	–30	90	10
59	18	–2	–36	72	28
59.5″–60–60.5″	30	–1	–30	30	58
			–96		
60.5″–61–61.5″	42	0	0	0	100
62	35	1	35	35	135
62.5″–63–63.5″	28	2	56	112	163
63.5″–64–64.5″	16	3	48	144	179
65	8	4	32	128	187
	187		171	611	
			+75		

$$\text{Mean} = 61 + \frac{75}{187} = 61.4, \text{ Mode} = 60.5 + \frac{35}{65} = 61.04$$

$$\sigma = \sqrt{\frac{611}{187} - \left(\frac{75}{187}\right)^2} = \sqrt{3.27 - 0.16} = \sqrt{3.11} = 1.76$$

$$\text{Skewness} = 61.4 - 61.04 = 0.36 \text{ inches.}$$

$$\text{Coefficient of skewness} = \frac{0.36}{1.76} = 0.205$$

Alternatively, we can determine the median as follows:

$$\text{Median} = \text{The size of } \frac{187}{2} \text{th item} = 93.5\text{th item}$$

$$= 60.5 + \frac{1 \times 35.5}{42} = 61.35$$

$$\text{Skewness} = 3(61.4 - 61.35) = 3(0.05) = 0.15$$

$$\text{Coefficient of skewness} = \frac{0.15}{1.76} = 0.09$$

The two coefficients are different because of the difficulties associated with determination of mode.

2. Bowley's (Quartile) measure of skewness

In the above two methods of measuring skewness, the whole series is taken into consideration. But, absolute as well as relative skewness may be secured even for a part of the series. The usual device is to measure the distance between the lower and the upper quartiles. In a symmetrical series, the quartiles would be equidistant from the value of the median, i.e.

$$\text{Median} - Q_1 = Q_3 - \text{Median}$$

NOTES

In other words, the value of the median is the mean of Q_1 and Q_3. In a skewed distribution, quartiles would not be equidistant from median unless the entire asymmetry is located at the extremes of the series. Bowley has suggested the following formula for measuring skewness, based on the preceding facts.

$$\text{Absolute SK} = (Q_3 - \text{Me}) - (\text{Me} - Q_1)$$
$$= Q_3 + Q_1 - 2\ \text{Me} \qquad (2.12)$$

If the quartiles are equidistant from the median, i.e. $(Q_3 - \text{Md}) = (\text{Md} - Q_1)$, then SK = 0. If the distance from the median to Q_1 exceeds that from Q_3 to the median, this will give a negative skewness. If the reverse is the case, it will give a positive skewness.

If the series expressed in different units are to be compared, it is essential to convert the absolute amount into the relative. Using the interquartile range as a denominator we have for the coefficient of skewness as follows:

$$\text{Relative SK} = \frac{Q_3 + Q_1 - 2\,\text{Md}}{Q_3 - Q_1} \qquad (2.13)$$

or,
$$\frac{(Q_3 - \text{Md}) - (\text{Md} - Q_1)}{(Q_3 - \text{Md}) + (\text{Md} - Q_1)}$$

If in the series the median and lower quartiles coincide, then the SK becomes (+1). If the median and upper quartiles coincide, then the SK becomes (–1).

This measure of skewness is rigidly defined and easily computable. Further, such a measure of skewness has the advantage that it has value limits between (+1) and (–1), with the result that it is sufficiently sensitive for many requirements. The only criticism levelled against such a measure is that it does not take into consideration all the items of these series, i.e. extreme items are neglected.

Example 2.35: Calculate the coefficient of skewness of the data of table given in Example 2.34 based on quartiles.

Solution: With reference to table given in Example 2.34, you have

$$Q_1 = \text{The size of } \frac{N}{4}\text{th}\left(= \frac{187}{4} = 46.75\text{th}\right) \text{ item}$$

$$= 59.5 + \frac{18.75}{30}$$
$$= 59.5 + 0.63 = 60.13$$

$$Q_3 = \text{The size of } \frac{3N}{4}\text{th item}\left(= \frac{3 \times 187}{4} = 140.25\text{th}\right) \text{ item}$$

$$= 62.5 + \frac{5.25}{28}$$
$$= 62.5 + 0.19 = 62.69$$
$$\text{Skewness} = 62.69 + 60.13 - 2\,(61.35) = 0.12$$

(using formula 2.12)

$$\text{Coefficient of skewness} = \frac{0.12}{62.69 - 60.13}$$

(using formula 2.13)

$$= \frac{0.12}{2.56} = 0.047$$

NOTES

3. Kelly's (Percentile) measure of skewness

To remove the defect of Bowley's measure that it does not take into account all the values, it can be enlarged by taking two deciles (or percentiles), equidistant from the median value. Kelly has suggested the following measure of skewness:

$$SK = P_{50} - \frac{P_{90} + P_{10}}{2}$$

$$= D_5 - \frac{D_9 + D_1}{2}$$

Though such a measure has got little practical use, yet theoretically this measure seems very sound.

Example 2.36: Calculate the Karl Pearson's coefficient of skewness from the following data:

Marks	*No. of Students*	*Marks*	*No. of Students*
above 0	150	above 50	70
" 10	140	" 60	30
" 20	100	" 70	14
" 30	80	" 80	0
" 40	80		

Solution:

Marks	*Frequency*	*Midpoint*	$X = (X - A)/10$	$f(X')$	$f(X'^2)$	*Cumulative Frequency (cf)*
0–10	10	5	–3	–30	90	10
10–20	40	15	–2	–80	160	50
20–30	20	25	–1	–20	20	70
				–130		
30–40	0	35	0	0	0	70
40–50	10	45	1	10	10	80
50–60	40	55	2	80	160	120
60–70	16	65	3	48	144	136
70–80	14	75	4	56	224	150
	150			194	808	
				+64		

Since it is a bimodal distribution Karl Pearson coefficient is appropriate and we need to calculate $\bar{X}$, Me and σ.

$$\bar{X} = 35 + \frac{64}{150} \times 10 = 35 + 4.27 = 39.27$$

NOTES

$$\text{Median} = \text{Size of } \frac{150}{2}\text{th item}$$

$$= 40 + \frac{10 \times 5}{10} = 45$$

$$\text{Standard deviation } (\sigma) = i \times \sqrt{\frac{\Sigma f(X'^2)}{N} - \left(\frac{\Sigma f(X')}{N}\right)^2}$$

$$= 10 \times \sqrt{\frac{808}{150} - \left(\frac{64}{150}\right)^2}$$

$$= 10 \times \sqrt{5.387 - 0.182}$$

$$= 10 \times 2.28 = 22.8$$

$$\text{Skewness} = \frac{3(\bar{X} - \text{Median})}{\sigma} = \frac{3(39.27 - 45)}{22.8}$$

$$= \frac{3(-5.73)}{22.8} = \frac{-17.19}{22.8} = -0.75$$

Example 2.37: From the following data compute quartile deviation and the coefficient of skewness.

Size	5–7	8–10	11–13	14–16	17–19
Frequency	14	24	38	20	4

Solution:

Size	*Frequency*	*Cumulative Frequency*
4.5–7.5	14	14
7.5–10.5	24	38
10.5–13.5	38	76
13.5–16.5	20	96
16.5–19.5	4	100

$$Q_1 = 7.5 + \frac{3 \times 11}{24} = 8.87$$

$$Q_3 = 10.5 + \frac{3 \times 37}{38} = 10.5 + \frac{111}{38} = 10.5 + 2.92 = 13.42$$

$$\text{Median} = 10.5 + \frac{3 \times 12}{38} = 10.5 + \frac{36}{38} = 10.5 + 0.947 = 11.447$$

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2} = \frac{13.42 - 8.87}{2} = \frac{4.55}{2} = 2.275$$

$$\text{Skewness} = \frac{Q_3 + Q_1 - 2\text{Me}}{Q_3 - Q_1}$$

$$= \frac{13.42 + 8.87 - 22.89}{13.42 - 8.87}$$

$$= \frac{-0.6}{4.55} = -0.13$$

Example 2.38: In a certain distribution the following results were obtained:

$\bar{X}$ = 45.00; Median = 48.00

Coefficient of Skewness = – 0.4

You are required to estimate the value of standard deviation.

Solution:

$$\text{Skewness} = \frac{3\,(\text{Mean} - \text{Median})}{\sigma}$$

$$-0.4 = \frac{3\,(45 - 48)}{\sigma}$$

$$-0.4\sigma = -9$$

$$\sigma = \frac{9}{0.4} = 22.5$$

Example 2.39: Karl Pearson's coefficient of skewness of a distribution is +0.32. Its standard deviation is 6.5 and mean is 29.6. Find the mode and median of the distribution.

Solution:

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$0.32 = \frac{29.6 - \text{Mode}}{6.5}$$

or $$6.5 \times 0.32 = 29.6 - \text{Mode}$$

$$\text{Mode} = 29.6 - 2.08 = 27.52$$

$$\text{Coefficient of skewness} = \frac{3\,(\text{Mean} - \text{Median})}{\sigma}$$

$$0.32 = \frac{3(29.6 - \text{Median})}{6.5}$$

$$6.5 \times 0.32 = 88.8 - 3\ \text{Median}$$

$$\text{Median} = \frac{88.8 - 2.08}{3} = 28.91$$

Example 2.40: You are given the position in a factory before and after the settlement of an industrial dispute. Comment on the gains or losses from the point of the workers and that of the management.

NOTES

	Before	*After*
No. of Workers	2440	2359
Mean Wages	45.5	47.5
Median Wages	49.0	45.0
Standard Deviation	12.0	10.0

Solution:

Employment. Since the number of workers employed after the settlement is less than the number of employed before, it has gone against the interest of the workers.

Wages. The total wages paid after the settlement were 2350 × 47.5 = Rs 1,11,625; before the settlement the amount disbursed was 2400 × 45.5 = Rs 1,09,200.

This means that the workers as a group are better off now than before the settlement, and unless the productivity of workers has gone up, this may be against the interest of management.

Uniformity in the wage structure. The extent of relative uniformity in the wage structure before and after the settlement can be determined by a comparison of the coefficient of variation.

$$\text{Coefficient of variation, before} = \frac{12}{45.5} \times 100 = 26.4$$

$$\text{Coefficient of variation, after} = \frac{10}{47.5} \times 100 = 21.05$$

This clearly means that there is comparatively lesser disparity in due wages received by the workers. Such a position is good for both the workers and the management.

Pattern of the wage structure. A comparison of the mean with the median leads to the obvious conclusion that before the settlement more than 50 per cent of the workers were getting a wage higher than this mean, i.e. (Rs. 45.5). After the settlement the number of workers whose wages were more than Rs. 45.5 became less than 50 per cent. This means that the settlement has not been beneficial to all the workers. It is only 50 per cent workers who have been benefited as a result of an increase in the total wages bill.

CHECK YOUR PROGRESS

14. What is skewness?
15. Name the measures of skewnness.

2.7 SUMMARY

In this unit, you have learned that:

- The most important objective of statistical analysis is to determine a single value for the entire mass of data so that it describes the overall level of the

group of observations and can be considered a representative value of the whole set of data.

- Variance or coefficient of variation has the same properties as standard deviation and is the square of standard deviation, represented as σ^2.
- The measure of dispersion is an important property of a distribution and needs to be measured by appropriate statistics.
- The measure of dispersion describes absolute and relative measures of dispersion. In its absolute form, it states the actual amount by which the value of an item on average deviates from a measure of central tendency.
- Absolute measures are expressed in concrete units, i.e. units in terms of which the data have been expressed, e.g. rupees, centimetres, kilograms, etc., and are used to describe frequency distribution.
- A relative measure of dispersion is a quotient obtained by dividing the absolute measures by a quantity in respect of which absolute deviation has been computed. It is as such a pure number and is usually expressed in a percentage form.
- Relative measures are used for making comparisons between two or more distributions.
- Skewness refers to lack of symmetry in distribution.

NOTES

2.8 KEY TERMS

- **Mean:** It is basically an arithmetic average and is the measure of central location.
- **Mode:** It is also a form of average and can be defined as the most frequently occurring value in the data.
- **Median:** It is a measure of central tendency and it appears in the centre of an ordered data.
- **Quartile:** It divides the data into four equal parts.
- **Deciles:** It divides the total ordered data into ten equal parts.
- **Percentile:** It divides the data into hundred equal parts.
- **Variance:** It is the square of standard deviation, represented as σ^2, and has the same properties as standard deviation.
- **Range:** It is the difference between the maximum and minimum values of a set of numbers. It indicates the limits within which the values fall.
- **Quartile deviation:** This deviation refers to a type of range based on the quartiles.
- **Mean deviation:** It is the arithmetic mean of the absolute deviations of a series of values.

NOTES

- **Standard deviation:** It is the measure of the dispersion of a set of values and is calculated from the mean of squared deviations.
- **Skewness:** It is a measure of asymmetry of the distribution of a real-valued random variable. A skew can be positive or negative.

2.9 ANSWERS TO 'CHECK YOUR PROGRESS'

1. A variable is considered as continuous when it can assume an infinite number of real values and it is considered discrete when it is the finite number of real values.
2. When the interest of the investigator focuses on the number of items below a specified value, then this specified value is the upper limit of the class interval. It is known as less than cumulative frequency distribution.
3. A single number describing some feature of a frequency distribution is called descriptive statistics.
4. The arithmetic mean is also commonly known as simply the mean. Even though average, in general, means any measure of central location. When we use the word average in our daily routine, we always mean the arithmetic average.
5. The advantages of mean are as follows:
   - It is a measure that can be easily calculated.
   - It includes all values of the data set in its calculation.
   - Its value varies very little from sample to sample taken from the same population.
   - It is useful for performing statistical procedures such as computing and comparing the means of several data sets.
6. Mean = 68

   Mode = 80

   Median = 72
7. Mean = 7
8. Rs 162.1
9. 98
10. 36.6 approx. and 22.6 approx.
11. The absolute measure of dispersion states the actual amount by which an item on average deviates from a measure of central tendency.
12. The relative measure of dispersion is a quotient computed by dividing the absolute measures by a quantity in respect of which the absolute deviation has been computed.

13. The range of a set of numbers is the difference between the maximum and minimum values. It indicates the limits within which the values fall.
14. Skewness refers to the lack of symmetry in a distribution. In the symmetrical distribution, the mean, median and mode coincide.
15. The following are the measures of skewness:
    - Relationship between three measures of central tendency—commonly known as the Karl Pearson's measure of skewness
    - Quartile measure of skewness—known as Bowley's measure of skewness
    - Percentile measure of skewness—also called the Kelly's measure of skewness
    - Measures of skewness based on moments

2.10 QUESTIONS AND EXERCISES

Short-Answer Questions

1. What do you understand by frequency distribution?
2. How would you convert discrete data into continuous data?
3. What do you mean by descriptive statistics?
4. How is central tendency measured?
5. Define the term arithmetic mean.
6. Write three characteristics of mean.
7. What is the importance of arithmetic mean in statistics?
8. Explain the term median with example.
9. How is location of median calculated using graphic analysis?
10. Define quartiles, deciles and percentiles with suitable examples.
11. What is the coefficient of variation?
12. What is range? How is it measured?
13. Write the definition and formula of quartile deviation.
14. How will you calculate the mean deviation of given data?
15. Explain standard deviation. Why is it used in statistical evaluation of data?
16. Calculate standard deviation for the series 1, 2, 3, 5, 7.
17. For a group of 50 male workers the mean and standard deviation of their weekly wages are Rs 63 and Rs 9 respectively. For a group of 40 female workers these are Rs 54 and Rs 6 respectively. Find the standard deviation of the combined group of 90 workers.

18. (a) Mean and standard deviations of two distributions of 100 and 150 items are 50, 5 and 40, 6 respectively. Find the mean and standard deviations of all the 250 items taken together.
    (b) Mean and standard deviations of 100 items are found by a student as 9 and if at the time of calculations two items are wrongly taken as 40 and 50 instead of 60 and 30, find the correct mean and standard deviations.
19. How will you measure skewness?

Long-Answer Questions

1. Explain the guidelines to be considered in constructing a frequency distribution.
2. Define the various measures of central tendency. What purposes do their measurement serve?
3. Define geometric and harmonic mean and explain their uses.
4. Show the relative positions of different averages in a moderately symmetrical series.
5. What do you mean by the following:
   (a) Quartiles
   (b) Deciles
   (c) Percentiles
6. What are the qualities which an average must possess? Which of the averages possess most of these qualities?
7. What do you mean by 'weights'? Why are they assigned? Point out a few cases in which weighted average should be used.
8. Differentiate between crude and corrected death rates.
9. The expenditure of ten families in rupees are given below:

Family	A	B	C	D	E	F	G	H	I	J
Expenditure	30	70	10	75	500	8	52	250	50	36

   Calculate the arithmetic average by (a) direct method and (b) short-cut method.
10. Calculate mean deviation and its coefficient about median, arithmetic mean and mode for the following figures, and show that the mean deviation about the median is the least.

    103, 50, 68, 110, 108, 105, 174, 103, 150, 200, 225, 350, 103

11. Compute mean deviations of the two series and point out which is more variable.

Month	*Index No. Calcutta*	*Index No. Delhi*	*Month*	*Index No. Calcutta*	*Index No. Delhi*
1970 April	93	107	1970 October	97	107
1970 May	97	108	1970 November	97	105
1970 June	95	102	1970 December	92	101
1970 July	95	102	1971 January	93	100
1970 August	95	102	1971 February	89	97
1970 September	95	104	1971 March	89	96

12. Calculate (a) median coefficient of dispersion and (b) mean coefficient of dispersion from the following data:

Size of Items	14	16	18	20	22	24	26
Frequency	2	4	5	3	2	1	4

13. Compute the mean deviation from the median and from the mean for the following distribution of the scores of 50 college students. Also complete the class interval.

Scores	140–	150–	160–	170–	180–	190–200
Frequency	4	6	10	18	9	3

14. Find the mean deviation about the mean of the following data of ages of married men in a certain town.

Ages	15–24	25–34	35–44	45–54	55–64	65–74
No. of Men	33	264	303	214	128	58

15. Calculate the mean deviation from the following data. What light does it throw on the social conditions of the community?

Difference in age between husband and wife:

Difference in Years:	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40
Frequency:	449	705	507	281	109	52	16	4

16. The following figures give the income of 10 persons in rupees. Find the standard deviation.

114, 115, 123, 120, 110, 130, 119, 118, 116, 115

17. Calculate the mean and standard deviation of the following values of the world's annual gold output in millions of pound (for 20 different years)

94, 95, 96, 93, 87, 79, 73, 69, 68, 67, 78, 82, 83, 89, 95, 103, 108, 117, 130, 97.

Also calculate the percentage of cases lying outside the mean at distances $\pm\sigma$, $\pm 2\sigma$, $\pm 3\sigma$ where σ denotes the standard deviation.

NOTES

18. Calculate standard deviation from the following data:

Size of Item	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

19. Calculate the arithmetic mean and the standard deviation from the following data:

Class Interval	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45
Frequency	6	5	15	10	5	4	3	2

20. Calculate the mean and the standard deviations from the following data:

Age Group Below	20–25	25–30	30–35	35–40	40–45	45–50	50–55	55 and above
No. of Employees	26	44	60	101	109	846	66	10

21. Calculate mean and standard deviation from the following data:

Age Under	10	20	30	40	50	60	70	80
No. of Persons Dying	15	30	53	75	100	110	115	125

22. The marks obtained by the students of class *A* and *B* are given below:

Marks	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45
Class *A*	1	10	20	8	6	3	1	–
Class *B*	5	6	15	10	5	4	2	2

Calculate mean, median, mode and standard deviation for the distributions. Explain your results regarding composition of the class in respect of intelligence.

23. Explain clearly the ideas implied in using arbitrary working origin and scale for the calculation of the arithmetic mean and standard deviation of frequency distribution.

The values of arithmetic mean and standard deviation of the following frequency distribution of a continuous variable derived from analysis are Rs 135.33 and Rs 9.6 respectively. Find the upper and lower limits of the various classes:

X'	–4	–3	–2	–1	0	+1	+2	+3
f	2	5	8	18	22	13	8	4

24. (a) Mean of 100 items is 50 and their standard deviation is 4. Find the sum and sum of squares of all the items.

(b) The mean and the standard deviation of a sample of 100 observations were calculated as 40 and 5.1, respectively, by a student, who took by mistake 50 instead of 40 for one observation. Calculate the correct mean and standard deviation.

25. The following data give the arithmetic averages and standard deviations of the three subgroups. Calculate the arithmetic average and standard deviation of the whole group.

Subgroup	*No. of Men*	*Average Wage (Rs)*	*SD of Wage (Rs)*
A	50	61.0	8.0
B	100	70.0	9.0
C	120	80.5	10.0

26. For a group containing 100 observations, the arithmetic mean and standard deviation are 8 and $\sqrt{10.5}$ respectively. For 50 observations selected from these 100 observations the mean and the standard deviation are 10 and 2 respectively. Find the arithmetic mean and the standard deviation of the other half.
27. A group has $\sigma = 10$, $N = 60$, $\sigma^2 = 4$. A subgroup of this has $\overline{X}_1 = 11$, $N_1 = 40$, $\sigma_1^2 = 2.25$. Find the mean and standard deviation of the other subgroup.
28. Two cricketers scored the following runs in the several innings. Find who is a better run-getter and who is a more consistent player.

    A 42, 17, 83, 59, 72, 76, 64, 45, 40, 32

    B 28, 70, 31, 0, 59, 108, 82, 14, 3, 95
29. The following are some of the particulars of the distribution of weights of boys and girls in a class.

	Boys	*Girls*
Number	100	50
Mean weight (kg)	60	45
Variance	9	4

    (a) Find standard deviation of the combined data.

    (b) Which of the two distributions is more variable?

2.11 FURTHER READING

Anderson, David R., Dennis J. Sweeney and Thomas A. Williams. *Essentials of Statistics for Business and Economics*. Mumbai: Thomson Learning, 2007.

Anderson, David R., Dennis J. Sweeney and Thomas A. Williams. *Quantitative Methods for Business*. Mumbai: Thomson Learning, 2005.

Bhardwaj, R.S. *Business Statistics*. New Delhi: Excel Books, 2000.

Chandan, J.S. *Business Statistics.* New Delhi: Vikas Publishing House, 2004.

Gupta, C.B. and Vijay Gupta. *An Introduction to Statistical Methods*. New Delhi: Vikas Publishing House, 2004.

NOTES

Hooda. R.P. *Statistics for Business & Economics.* New Delhi: Macmillan India Ltd., 2004.

Kothari C.R. *Quantitative Techniques*. New Delhi: Vikas Publishing House, 1984.

Levin, Richard I. and David S. Rubin. *Statistics for Business*. New Delhi: Prentice Hall of India, 1990.

Monga, G.S. *Mathematics and Statistics for Economics*. New Delhi: Vikas Publishing House.

Sancheti D.C. and V.K. Kapoor. *Business Mathematics*. New Delhi: Sultan Chand & Sons.

Zameeruddin Qazi, V.K. Sharma and S.K. Bhambri. *Business Mathematics*. New Delhi: Vikas Publishing House, 2008.

UNIT 3 CORRELATION AND REGRESSION ANALYSES

Structure

3.0 INTRODUCTION

In this unit, you will learn about correlation and regression analyses. Correlation analysis looks at the indirect relationships in sample survey data and establishes the variables which are most closely associated with a given action or mindset. It is the process of finding how accurately the line fits using the observations. Correlation analysis can be referred to as the statistical tool used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. In fact, the word correlation refers to the relationship or interdependence between two variables. There are various phenomena which have relation to one another. The theory by means of which quantitative connections between two sets of phenomena are determined is called the 'theory of correlation'. On the basis of the theory of correlation, you can study the comparative changes occurring in two related phenomena and their cause–

NOTES

effect relation can also be examined. Thus, correlation is concerned with relationship between two related and quantifiable variables and can be positive or negative.

Regression analysis is the mathematical process of using observations to find the *line of best fit* through the data in order to make estimates and predictions about the behaviour of the variables. This technique is used to determine the statistical relationship between two or more variables and to make prediction of one variable on the basis of one or more other variables. While making use of the regression techniques for making predictions, it is always assumed that there is an actual relationship between the dependent and independent variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable. You will also learn about the scatter diagram, least squares method and standard error of estimate. Standard error of estimate is a measure developed by the statisticians for measuring the reliability of the estimating equation. The larger the standard error of estimate (SE*e*), the greater the dispersion, or scattering, of given observations around the regression line. But if the SE of estimate happens to be zero, then the estimating equation is a 'perfect' estimator, i.e. a cent per cent correct estimator of the dependent variable. You will be able to interpret the coefficient of determination, i.e. r^2, using the coefficient of correlation.

In this unit, you will also learn about the time series analysis and its importance in making forecasts and predictions. A time series is a set of ordered observations on a quantitative characteristic of a phenomenon at equally spaced time points. One of the main goals of time series analysis is to forecast future values of the series on the basis of the trend which is a regular, slowly evolving change in the series level. Time series can be defined as 'a set of numeric observations of the dependent variable, measured at specific points in time in chronological order, usually at equal intervals, in order to determine the relationship of time to such variable'.

3.1 UNIT OBJECTIVES

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

- Explain correlation analysis
- Evaluate coefficient of determination and coefficient of correlation
- Interpret coefficient of determination, r^2
- Calculate probable error of the coefficient of correlation
- Calculate correlation using various methods
- Understand the regression analysis
- Describe how are assumptions made in regression analysis
- Explain simple linear regression model
- Define scatter diagram method and least squares method

- Understand the relationship between correlation and regression coefficient
- Explain the influences of time series analysis
- Understand the techniques used in trend analysis

NOTES

3.2 CORRELATION ANALYSIS

Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable. In fact, the word 'correlation' refers to the relationship or interdependence between two variables. There are various phenomenons which have relation to each other. For instance, when demand of a certain commodity increases, its price goes up and when its demand decreases, its price comes down. Similarly, with age the height of the children goes up, with height the weight of the children goes up and with money the supply and the general level of prices go up. Such sort of relationship can also well be noticed for several other phenomena. The theory by means of which quantitative connections between two sets of phenomena are determined is called the '*Theory of Correlation*'.

On the basis of the theory of correlation, one can study the comparative changes occurring in two related phenomena and their cause-effect relation can be examined. It should, however, be borne in mind that relationship like 'black cat causes bad luck', 'filled up pitchers result in good fortune' and similar other beliefs of the people cannot be explained by the theory of correlation, since they are all imaginary and are incapable of being justified mathematically. Thus, correlation is concerned with relationship between two related and quantifiable variables. If two quantities vary in sympathy so that a movement (an increase or decrease) in the one tends to be accompanied by a movement in the same or opposite direction in the other and the greater the change in the one, the greater is the change in the other, the quantities are said to be correlated. This type of relationship is known as correlation or what is sometimes called, in statistics, as covariation.

For correlation, it is essential that the two phenomena should have cause-effect relationship. If such relationship does not exist, then one should not talk of correlation. For example, if the height of the students as well as the height of the trees increase, then one should not call it a case of correlation because the two phenomena, viz., the height of students and the height of trees are not even casually related. But the relationship between the price of a commodity and its demand, the price of a commodity and its supply, the rate of interest and savings, etc. are the examples of correlation, since in all such cases the change in one phenomen is explained by a change in other phenomenon.

It is appropriate here to mention that correlation in case of phenomena pertaining to natural sciences can be reduced to absolute mathematical term, e.g. heat always increases with light. But in, phenomena pertaining to social sciences, it

NOTES

is often difficult to establish any absolute relationship between two phenomena. Hence, in social sciences we must take the fact of correlation being established if in a large number of cases, two variables always tend to move in the same or opposite direction.

Correlation can either be positive or negative. Whether correlation is positive or negative would depend upon the direction in which the variables are moving. If both variables are changing in the same direction, then correlation is said to be positive, but when the variations in the two variables take place in the opposite direction, the correlation is termed as negative. This can be explained as under (see Table 3.1).

Table 3.1 *Changes in Variables and Nature of Correlation*

Changes in Independent Variable	*Changes in Dependent Variable*	*Nature of Correlation*
Increase (+)↑	Increase (+)↑	Positive (+)
Decrease (–)↓	Decrease (–)↓	Positive (+)
Increase (+)↑	Decrease (–)↓	Negative (–)
Decrease (–)↓	Increase (+)↑	Negative (–)

Statisticians have developed *two measures for describing the correlation* between two variables, namely the coefficient of determination and the coefficient of correlation.

3.2.1 The Coefficient of Determination

The coefficient of determination (symbolically indicated as r^2, though some people would prefer to put it as R^2) is a measure of the degree of linear association or correlation between two variables, say *X* and *Y*, one of which happens to be independent variable and the other being dependent variable. This coefficient is based on the following two kinds of variations:

(*i*) The variation of the *Y* values around the fitted regression line, namely $\sum\left(Y-\hat{Y}\right)^2$, technically known as the unexplained variation.

(*ii*) The variation of the *Y* values around their own mean, namely $\sum\left(Y-\bar{Y}\right)^2$, technically known as the total variation.

If you subtract the unexplained variation from the total variation, you will obtain what is known as the explained variation, i.e. the variation explained by the line of regression. Thus,

$$\text{Explained variation} = (\text{Total variation}) - (\text{Unexplained variation})$$

$$= \sum\left(Y-\bar{Y}\right)^2 - \sum\left(Y-\hat{Y}\right)^2$$

$$= \sum\left(\hat{Y}-\bar{Y}\right)^2$$

NOTES

The total, explained and unexplained variations are shown in Figure 3.1.

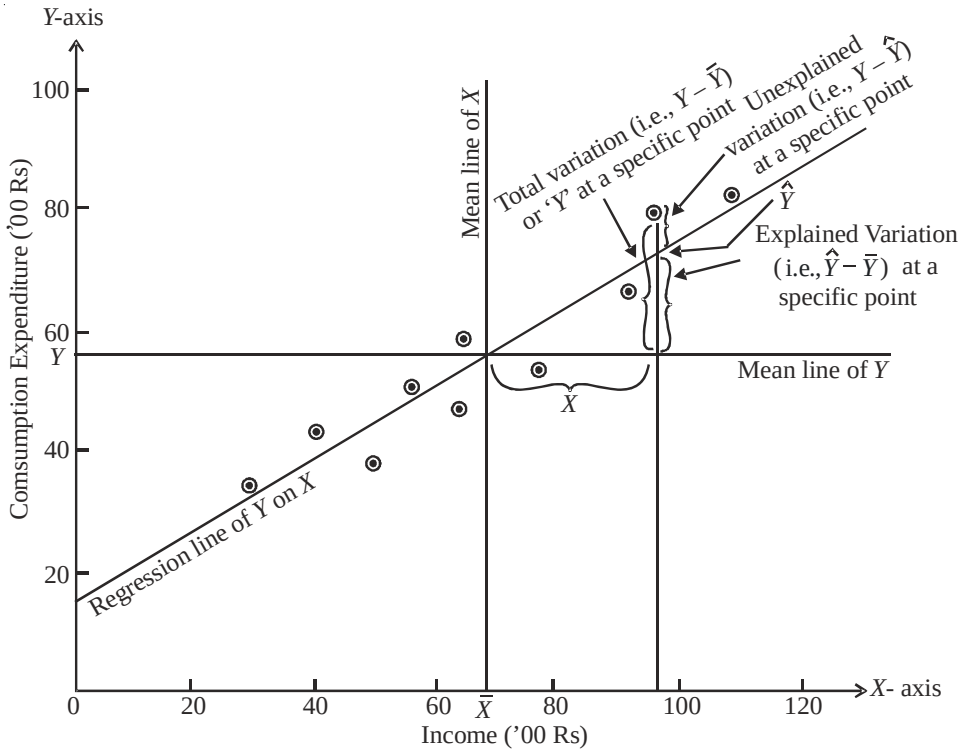


Fig. 3.1 *Diagram Showing Total, Explained and Unexplained Variations*

The coefficient of determination is that fraction of the total variation of *Y* which is explained by the regression line. In other words, coefficient of determination is the ratio of explained variation to total variation in the *Y* variable related to the *X* variable. Coefficient of determination can algebraically be stated as follows:

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= \frac{\Sigma\left(\hat{Y} - \bar{Y}\right)^2}{\Sigma\left(Y - \bar{Y}\right)^2}$$

Alternatively r^2 can also be stated as under:

$$r^2 = 1 - \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= 1 - \frac{\Sigma\left(\hat{Y} - \bar{Y}\right)^2}{\Sigma\left(Y - \bar{Y}\right)^2}$$

Interpreting r^2

The coefficient of determination can have a value ranging from zero to one. A value of one can occur only if the unexplained variation is zero which simply

NOTES

means that all the data points in the scatter diagram fall exactly on the regression line. For a zero value to occur, $\Sigma(Y-\bar{Y})^2 = \Sigma(Y-\hat{Y})^2$, which simply means that X tells us nothing about Y and hence there is no regression relationship between X and Y variables. Values between 0 and 1 indicate the 'Goodness of fit' of the regression line to the sample data. The higher the value of r^2, the better the fit. In other words, the value of r^2 will lie somewhere between 0 and 1. If r^2 has a zero value, then it indicates no correlation, but if it has a value equal to 1, then it indicates that there is perfect correlation and as such the regression line is a perfect estimator. But in most cases, the value of r^2 will lie somewhere between these two extremes of 1 and 0. You should remember that an r^2 close to 1 indicates a strong correlation between X and Y, while an r^2 near zero means there is little correlation between these two variables.

r^2 value can as well be interpreted by looking at the amount of the variation in Y, the dependant variable, that is explained by the regression line. Suppose you get a value of $r^2 = 0.925$. This would mean that the variations in independent variable (say X) would explain 92.5 per cent of the variation in the dependent variable (say Y). If r^2 is close to 1, then it indicates that the regression equation explains most of the variations in the dependent variable.

Example 3.1: Calculate the coefficient of determination (r^2) using data given in Example 3.8. Analyse the result.

Solution: r^2 can be worked out as shown below:

Since,
$$r^2 = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$
$$= 1 - \frac{\Sigma\left(Y-\hat{Y}\right)^2}{\Sigma\left(Y-\bar{Y}\right)^2}$$

As, $\Sigma\left(Y-\bar{Y}\right)^2 = \Sigma Y^2 = \left(\Sigma Y^2 - n\bar{Y}^2\right)$, we can write,

$$r^2 = 1 - \frac{\Sigma\left(Y-\hat{Y}\right)^2}{\Sigma Y^2 - n\bar{Y}^2}$$

Calculating and putting the various values, we have the following equation,

$$r^2 = 1 - \frac{260.54}{34223 - 10(56.3)^2} = 1 - \frac{260.54}{2526.10} = 0.897$$

Analysis of result: The regression equation used to calculate the value of coefficient of determination (r^2) from the sample data shows that about 90 per cent of the variations in consumption expenditure can be explained. In other words, it means that the variations in income explain about 90 per cent of variations in consumption expenditure.

3.2.2 Coefficient of Correlation

The coefficient of correlation, symbolically denoted by '*r*', is another important measure to describe how well one variable is explained by another. It measures the degree of relationship between the two casually related variables. The value of this coefficient can never be more than +1 or less than –1. Thus +1 and –1 are the limits of this coefficient. For a unit change in independent variable, if there happens to be a constant change in the dependent variable in the same direction, then the value of the coefficient will be +1 indicative of the perfect positive correlation; but if such a change occurs in the opposite direction, the value of the coefficient will be –1, indicating the perfect negative correlation. In practical life the possibility of obtaining either a perfect positive or perfect negative correlation is very remote particularly in respect of phenomena concerning social sciences. If the coefficient of correlation has a zero value, then it means that there exists no correlation between the variables under study.

There are several methods of finding the coefficient of correlation but the following three methods are considered important:

(*i*) Determination of Coefficient of Correlation by the Method of Least Squares

(*ii*) Determination of Coefficient of Correlation using Simple Regression Coefficients

(*iii*) Determination of Coefficient of Correlation through Product Moment Method or Karl Pearson's Coefficient of Correlation

Whichever of these above-mentioned three methods you adopt, you get the same value of *r*.

(*i*) Determination of coefficient of correlation by the method of least squares

Under this method, first of all the estimating equation is obtained using least square method of simple regression analysis. The equation is worked out as follows:

$$\hat{Y} = a + bX_i$$

Total variation $= \sum\left(Y - \overline{Y}\right)^2$

Unexplained variation $= \sum\left(Y - \hat{Y}\right)^2$

Explained variation $= \sum\left(\hat{Y} - \overline{Y}\right)^2$

Then by applying the following formulae you can find the value of the coefficient of correlation:

$$r = \sqrt{r^2} = \sqrt{\frac{\text{Explained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 - \frac{\text{Unexplained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}}$$

NOTES

This clearly shows that coefficient of correlation happens to be the square root of the coefficient of determination.

Short-cut formula for finding the value of '*r*' by the method of least squares can be repeated and readily written as follows:

$$r = \sqrt{\frac{a\Sigma Y + b\Sigma XY - n\bar{Y}^2}{\Sigma Y^2 - n\bar{Y}^2}}$$

Where,

a = *Y*-intercept

b = Slope of the estimating equation

X = Values of the independent variable

Y = Values of dependent variable

$\bar{Y}$ = Mean of the observed values of *Y*

n = Number of items in the sample (i.e. pairs of observed data)

The plus (+) or the minus (–) sign of the coefficient of correlation worked out by the method of least squares is related to the sign of '*b*' in the estimating equation, viz. $\hat{Y} = a + bX_i$. If '*b*' has a minus sign, the sign of '*r*' will also be minus but if '*b*' has a plus sign, the sign of '*r*' will also be plus. The value of '*r*' indicates the degree along with the direction of the relationship between the two variables *X* and *Y*.

(*ii*) Determination of coefficient of correlation using simple regression coefficients

Under this method, the estimating equation of *Y* and the estimating equation of *X* is worked out using the method of least squares. From these estimating equations, you find the regression coefficient of *X* on *Y*, i.e. the slope of the estimating equation of *X* (symbolically written as b_{XY}) and this happens to be equal to $r\frac{\sigma_X}{\sigma_Y}$, and similarly, you find the regression coefficient of *Y* on *X*, i.e. the slope of the estimating equation of *Y* (symbolically written as b_{YX}) and this happens to be equal to $r\frac{\sigma_Y}{\sigma_X}$. For finding '*r*', the square root of the product of these two regression coefficients are worked out as stated below:[1]

[1] Remember the short-cut formulae to workout b_{XY} and b_{YX}:

$$b_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2}$$

$$b_{YX} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

$$r = \sqrt{b_{XY} b_{YX}}$$

$$= \sqrt{r \frac{\sigma_X}{\sigma_Y} r \frac{\sigma_Y}{\sigma_X}}$$

$$= \sqrt{r^2}$$

$$= r$$

The sign of '*r*' will depend upon the sign of the regression coefficients. If they have minus sign, then '*r*' will take minus sign but the sign, of '*r*' will be plus if regression coefficients have plus sign.

3.2.3 Karl Pearson's Coefficient

The Karl Pearson's method is most widely used method of measuring the relationship between two variables. This coefficient is based on the following three assumptions:

(*i*) There is a linear relationship between the two variables, which means that straight line would be obtained if the observed data are plotted on a graph.

(*ii*) The two variables are casually related, which means that one of the variables is independent and the other one is dependent.

(*iii*) A large number of independent causes are operating in both the variables so as to produce a normal distribution.

According to Karl Pearson, '*r*' can be worked out as follows:

$$r = \frac{\sum XY}{n \sigma_X \sigma_Y}$$

where,

$X = (X - \bar{X})$

$Y = (Y - \bar{Y})$

σ_X = Standard deviation of X series and is equal to $\sqrt{\frac{\sum X^2}{n}}$

σ_Y = Standard deviation of Y series and is equal to $\sqrt{\frac{\sum Y^2}{n}}$

n = Number of pairs of X and Y observed.

A short-cut formula known as the Product Moment Formula can be derived from the preceding formula as follows:

$$r = \frac{\sum XY}{n \sigma_X \sigma_Y} = \frac{\sum XY}{\sqrt{\frac{\sum X^2}{n} \cdot \frac{\sum Y^2}{n}}}$$

$$= \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

NOTES

The above formulae are based on obtaining true means (namely $\bar{X}$ and $\bar{Y}$) first and then doing all other calculations. This happens to be a tedious task, particularly if the true means are in fractions. To avoid difficult calculations, you need to make use of the assumed means in taking out deviations and doing the related calculations. In such a situation, you can use the following formula for finding the value of '*r*':[2]

(*a*) *In case of ungrouped data:*

$$r = \frac{\frac{\sum dX.dY}{n} - \left(\frac{\sum dX}{n} \cdot \frac{\sum dY}{n}\right)}{\sqrt{\frac{\sum dX^2}{n} - \left(\frac{\sum dX}{n}\right)^2} \cdot \sqrt{\frac{\sum dY^2}{n} - \left(\frac{\sum dY}{n}\right)^2}}$$

$$= \frac{\sum dX.dY - \left(\frac{\sum dX \times \sum dY}{n}\right)}{\sqrt{\sum dX^2 - \frac{(\sum dX)^2}{n}} \sqrt{\sum dY^2 - \frac{(\sum dY)^2}{n}}}$$

where $\sum dX = \sum(X - X_A)$ X_A = Assumed average of X

$\sum dY = \sum(Y - Y_A)$ Y_A = Assumed average of Y

$\sum dX^2 = \sum(X - X_A)^2$

$\sum dY^2 = \sum(Y - Y_A)^2$

$\sum dXdY = \sum(X - X_A)(Y - Y_A)$

n = Number of pairs of observations of X and Y

(*b*) *In case of grouped data:*

$$r = \frac{\frac{\sum fdX.dY}{n} - \left(\frac{\sum fdX}{n} \cdot \frac{\sum fdY}{n}\right)}{\sqrt{\frac{\sum fdX^2}{n} - \left(\frac{\sum fdX}{n}\right)^2} \sqrt{\frac{\sum fdY^2}{n} - \left(\frac{\sum fdY}{n}\right)^2}}$$

[2] In case you take assumed mean to be zero for variable *X* as for variable *Y*, then your formula will be as follows:

$$r = \frac{\frac{\sum XY}{n} - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right)}{\sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} \sqrt{\frac{\sum Y^2}{n} - \left(\frac{\sum Y}{n}\right)^2}}$$

or

$$r = \frac{\frac{\sum XY}{n} - \bar{X}\bar{Y}}{\sqrt{\frac{\sum X^2}{n} - \bar{X}^2} \sqrt{\frac{\sum Y^2}{n} - \bar{Y}^2}}$$

$$r = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{\sum X^2 - n\bar{X}^2} \sqrt{\sum Y^2 - n\bar{Y}^2}}$$

or
$$r = \frac{\sum fdXdY - \left(\frac{\sum fdX \sum fdY}{n}\right)}{\sqrt{\sum fdX^2 - \left(\frac{\sum fdX}{n}\right)^2}\sqrt{\sum fdY^2 - \left(\frac{\sum fdY}{n}\right)^2}}$$

Where,
$$\sum fdXdY = \sum f(X - X_A)(Y - Y_A)$$
$$\sum fdX = \sum f(X - X_A)$$
$$\sum fdY = \sum f(Y - Y_A)$$
$$\sum fdY^2 = \sum f(Y - Y_A)^2$$
$$\sum fdX^2 = \sum f(X - X_A)^2$$
$$n = \text{Number of pairs of observations of } X \text{ and } Y$$

3.2.4 Probable Error (PE) of the Coefficient of Correlation

Probable Error (PE) of *r* is very useful in interpreting the value of *r* and is worked out as under for Karl Pearson's coefficient of correlation:

$$\text{PE} = 0.6745\frac{1 - r^2}{\sqrt{n}}$$

If *r* is less than its PE, it is not at all significant. If *r* is more than PE, there is correlation. *If r is more than 6 times its PE and greater than ± 0.5, then it is considered significant.*

Example 3.2: From the following data calculate '*r*' between *X* and *Y* applying the following three methods:

(a) The method of least squares.

(b) The method based on regression coefficients.

(c) The product moment method of Karl Pearson.

Verify the obtained result of any one method with that of another.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Solution: Let us develop the following table for calculating the value of '*r*':

X	*Y*	X^2	Y^2	*XY*
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98

8	16	64	256	128
9	15	81	225	135
n=9				
$\sum X = 45$	$\sum Y = 108$	$\sum X^2 = 285$	$\sum Y^2 = 1356$	$\sum XY = 597$

NOTES

$\therefore \quad \bar{X} = 5; \qquad \bar{Y} = 12$

(*i*) *Coefficient of correlation by the method of least squares is worked out as under*:

First of all, find out the estimating equation,

$$\hat{Y} = a + bX_i$$

where $$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

$$= \frac{597 - 9(5)(12)}{285 - 9(25)} = \frac{597 - 540}{285 - 225} = \frac{57}{60} = 0.95$$

and $$a = \bar{Y} - b\bar{X}$$

$$= 12 - 0.95(5) = 12 - 4.75 = 7.25$$

Hence, $$\hat{Y} = 7.25 + 0.95X_i$$

Now '*r*' can be worked out as under by the method of least squares:

$$r = \sqrt{1 - \frac{\text{Unexplained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2}} = \sqrt{\frac{\sum\left(\hat{Y} - \bar{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2}}$$

$$= \sqrt{\frac{a\sum Y + b\sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}}$$

This is as per short-cut formula,

$$r = \sqrt{\frac{7.25(108) + 0.95(597) - 9(12)^2}{1356 - 9(12)^2}}$$

$$= \sqrt{\frac{783 + 567.15 - 1296}{1356 - 1296}}$$

$$= \sqrt{\frac{54.15}{60}} = \sqrt{0.9025} = 0.95$$

NOTES

(ii) Coefficient of correlation by the method based on regression coefficients is worked out as under:

Regression coefficients of *Y* on *X*,

$$\text{i.e.} \qquad b_{YX} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

$$= \frac{597 - 9 \times 5 \times 12}{285 - 9(5)^2} = \frac{597 - 540}{285 - 225} = \frac{57}{60}$$

Regression coefficient of *X* on *Y*,

$$\text{i.e.} \qquad b_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2}$$

$$= \frac{597 - 9 \times 5 \times 12}{1356 - 9(12)^2} = \frac{597 - 540}{1356 - 1296} = \frac{57}{60}$$

$$\text{Hence,} \qquad r = \sqrt{b_{YX} \cdot b_{XY}}$$

$$= \sqrt{\frac{57}{60} \times \frac{57}{60}} = \frac{57}{60} = 0.95$$

(iii) Coefficient of correlation by the product moment method of Karl Pearson is worked out as under:

$$r = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{\sum X^2 - n\bar{X}^2}\sqrt{\sum Y^2 - n\bar{Y}^2}}$$

$$= \frac{597 - 9(5)(12)}{\sqrt{285 - 9(5)^2}\sqrt{1356 - 9(12)^2}}$$

$$= \frac{597 - 540}{\sqrt{285 - 225}\sqrt{1356 - 1296}} = \frac{57}{\sqrt{60}\sqrt{60}} = \frac{57}{60} = 0.95$$

Hence, you get the value of $r = 0.95$. You get the same value applying the other two methods also. Therefore, whichever method you apply, the results will be the same.

Example 3.3: Calculate the coefficient of correlation and lines of regression from the following data and comment.

	X *Advertising Expenditure* (*Rs '00*)				
Y *Sales Revenue* (Rs '000)	5–15	15–25	25–35	35–45	Total
75–125	3	4	4	8	19
125–175	8	6	5	7	26
175–225	2	2	3	4	11
225–275	2	3	2	2	9
Total	15	15	14	21	$n = 65$

NOTES

Solution: Since the given information is a case of bivariate grouped data we shall extend the given table rightwards and downwards to obtain various values for finding '*r*' as stated here:

Y Sales Revenue (*Rs '000*)	*X* Advertising Expenditure (Rs '00) 5-15	15-25	25-35	35-40	*Total* (*f*)	*Midpoint* of *Y*	*If A*=200 *i*=50 ∴ *dY*	*fdY*	*fdY*²	*fdX.dY*
75-125	3	4	4	8	19	100	–2	–38	76	4
125-175	8	6	5	7	26	150	–1	–26	26	15
175-225	2	2	3	4	11	200	0	0	0	0
225-275	2	3	2	2	9	250	1	9	9	–5
Total (or *f*)	15	15	14	21	*n* = 65			Σ*fdY* = –55	Σ*fdY*² = 111	Σ*fdX.dY* = 14
Midpoint of *X*	10	20	30	40						
If *A* = 30 *i* =10 ∴ *dX*	–2	–1	0	1						
fdX	–30	–15	0	21	Σ*fdX* = –24					
*fdX*²	60	15	0	21	Σ*fdX*² = 96					
fdX.dY	24[3]	11	0	–21	Σ*fdX·dY* =14					

$$\therefore \quad r = \frac{\frac{\sum fdX.dY}{n} - \left(\frac{\sum fdX}{n}\frac{\sum fdY}{n}\right)}{\sqrt{\frac{\sum fdX^2}{n} - \left(\frac{\sum fdX}{n}\right)^2}\sqrt{\frac{\sum fdY^2}{n} - \left(\frac{\sum fdY}{n}\right)^2}}$$

Putting the calculated values in the above equation, you have:

$$r = \frac{\frac{14}{65} - \left(\frac{-24}{65} \times \frac{-55}{65}\right)}{\sqrt{\frac{96}{65} - \left(\frac{-24}{65}\right)^2}\sqrt{\frac{111}{65} - \left(\frac{-55}{65}\right)^2}}$$

[3] This value has been worked out as under:

f	*dXdY*	=	*fdX dY*
(3)	(–2) (–2)	=	12
(8)	(–2) (–1)	=	16
(2)	(–2) (0)	=	0
(2)	(–2) (1)	=	–4
	Total		24

Similarly, for other columns also, the *fdXdY* values can be obtained. The process can be repeated for finding *fdXdY* values row-wise and finally Σ*fdXdY* can be checked.

NOTES

$$= \frac{0.2154-(+0.3124)}{\sqrt{1.48-0.14}\sqrt{1.71-0.72}}$$

$$= \frac{(-)0.0970}{\sqrt{(1.34)\times(99)}} = \frac{-0.00970}{\sqrt{1.3266}} = \frac{-0.0970}{1.15} = (-)0.0843$$

Hence, $r = (-)0.0843$

This shows a poor negative correlation between the two variables. Since only 0.64 per cent [r^2 being $(0.08)^2 = 0.0064$] variation in Y (sales revenue) is explained by variation in X (advertising expenditure).

The two lines of regression are as under:

Regression line of X on Y: $\left(X-\overline{X}\right) = r\frac{\sigma_X}{\sigma_Y}\left(Y-\overline{Y}\right)$

Regression line of Y on X: $\left(Y-\overline{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X-\overline{X}\right)$

First obtain the following values:

$$\overline{X} = A + \frac{\sum fdX}{n} i = 30 + \frac{(-24)}{65}\times 10 = 26.30$$

$$\overline{Y} = A + \frac{\sum fdY}{n} i = 200 + \frac{-55}{65}\times 10 = 157.70$$

$$\sigma_X = \sqrt{\frac{\sum fdX^2}{n} - \left(\frac{\sum fdX}{n}\right)^2}\times i = \sqrt{\frac{96}{65} - \left(\frac{-24}{65}\right)^2}\times 10 = 11.60$$

$$\sigma_Y = \sqrt{\frac{\sum fdY^2}{n} - \left(\frac{\sum fdY}{n}\right)^2}\times i = \sqrt{\frac{111}{65} - \left(\frac{-55}{65}\right)^2}\times 50 = 49.50$$

Therefore, the regression line of X on Y:

$$(X-26.30) = \frac{11.6}{49.5}(-0.084)(Y-157.70)$$

or $$\hat{X} = -0.02Y + 3.15 + 26.30$$

$$\boxed{\therefore \hat{X} = -0.02Y + 29.45}$$

Regression line of Y on X:

$$(Y-157.70) = \frac{49.5}{11.6}(-0.084)(X-26.30)$$

or $$\hat{Y} = -0.36X + 9.47 + 157.70$$

$$\boxed{\therefore \hat{Y} = -0.36X + 167.17}$$

NOTES

3.2.5 Coefficients of Non-Determination and Alienation

Alongwith coefficients of determination and correlation, you need to learn two other coefficients: coefficient of non-determination and coefficient of alienation.

(*i*) *Coefficient of non-determination*. Instead of using coefficient of determination, sometimes coefficient of non-determination is used. Coefficient of non-determination (denoted by k^2) is the ratio of unexplained variation to total variation in the *Y* variable related to the *X* variable. Algebrically, you can write it as follows:

$$k^2 = \frac{\text{Unexplained variation}}{\text{Total variation}} = \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}$$

Concerning the data of Example 3.1, the coefficient of non-determination will be calculated as follows:

$$k^2 = \frac{260.54}{2526.10} = 0.103$$

The value of k^2 shows that about 10 per cent of the variation in consumption expenditure remains unexplained by the regression equation you had worked out, namely $\hat{Y} = 14.000 + 0.616X_i$. In simple terms, this means that variable other than *X* is responsible for 10 per cent of the variations in the dependent variable *Y* in the given case.

Coefficient of non-determination can also be worked out as under:

$$k^2 = 1 - r^2$$

Accordingly, for Example 3.1, it will be equal to 1–0.897 = 0.103

Note: *Always remember that* $r^2 + k^2 = 1$.

(*ii*) *Coefficient of alienation*. Based on k^2 you can work out one more measure, namely the coefficient of alienation, symbolically written as '*k*'.

Thus, coefficient of alienation, i.e. '*k*' $= \sqrt{k^2}$

Unlike $r + k^2 = 1$, the sum of '*r*' and '*k*' will not be equal to 1 unless one of the two coefficients is 1 and in this case the remaining coefficients must be zero. In all other cases, '*r*' + '*k*' > 1. Coefficient of alienation is not a popular measure from practical point of view and is used very rarely.

3.2.6 Spearman's Rank Correlation

If observations on two variables are given in the form of ranks and not as numerical values, it is possible to compute what is known as rank correlation between the two series.

The rank correlation, written ρ, is a descriptive index of agreement between ranks over individuals. It is the same as the ordinary coefficient of correlation computed on ranks, but its formula is simpler.

$$\rho = 1 - \frac{6\Sigma D_i^2}{n(n^2 - 1)}$$

Here, n is the number of observations and D_i the positive difference between ranks associated with the individuals i.

Like r, the rank correlation lies between –1 and +1.

NOTES

Example 3.4: The ranks given by two judges to 10 individuals are as follows:

	Rank given by			
Individual	*Judge I*	*Judge II*	*D*	D^2
	x	y	$= x - y$	
1	1	7	6	36
2	2	5	3	9
3	7	8	1	1
4	9	10	1	1
5	8	9	1	1
6	6	4	2	4
7	4	1	3	9
8	3	6	3	9
9	10	3	7	49
10	5	2	3	9
				$\Sigma D^2 = 128$

Determine the rank correlation.

Solution: The rank correlation is given by

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 128}{10^3 - 10} = 1 - 0.776 = 0.224$$

The value of $\rho = 0.224$ shows that the agreement between the judges is not high.

Example 3.5: Based on data given in Example 3.4, compute r and compare.

Solution: The simple coefficient of correlation r for the previous data is calculated as follows:

x	y	$□^2$	y^2	xy
1	7	1	49	7
2	5	4	25	10
7	8	49	64	56
9	10	81	100	90
8	9	64	81	72
6	4	36	16	24
4	1	16	1	4
3	6	9	36	18
10	3	100	9	30
5	2	25	4	10
$\Sigma x = 55$	$\Sigma y = 55$	$\Sigma x^2 = 385$	$\Sigma y^2 = 385$	$\Sigma xy = 321$

NOTES

$$r = \frac{321 - 10 \times \frac{55}{10} \times \frac{55}{10}}{\sqrt{385 - 10 \times \left(\frac{55}{10}\right)^2}\sqrt{385 - 10 \times \left(\frac{55}{10}\right)^2}} = \frac{18.5}{\sqrt{82.5 \times 82.5}} = \frac{18.5}{82.5} = 0.224$$

This shows that the Spearman ρ for any two sets of ranks is the same as the Pearson r for the set of ranks. However, it is much easier to compute ρ.

Often, the ranks are not given. Instead, the numerical values of observations are given. In such a case, you must attach the ranks to these values to calculate ρ.

Example 3.6:

Marks in Maths	*Marks in Stats*	*Rank in Maths*	*Rank in Stats*	*D*	D^2
45	60	4	2	2	4
47	61	3	1	2	4
60	58	1	3	2	4
38	48	5	4	1	1
50	46	2	5	3	9
					$\Sigma D^2 = 22$

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 22}{125 - 5} = -0.1$$

This shows a negative, though small, correlation between the ranks.

If two or more observations have the same value, their ranks are equal and obtained by calculating the means of the various ranks.

If in these data, marks in maths are 45 for each of the first two students, the rank of each would be $\frac{3+4}{2} = 3.5$. Similarly, if the marks of each of the last two students in statistics are 48, their ranks would be $\frac{4+5}{2} = 4.5$

The problem takes the following shape:

Marks in Maths	*Marks in Stats*	*Rank* *x*	*y*	*D*	D^2
45	60	3.5	2	1.5	2.25
45	61	3.5	1	2.5	6.25
60	58	1	3	2	4.00
38	48	5	4.5	1.5	2.25
50	48	2	4.5	2.5	6.25

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 21}{120} = -0.05$$

NOTES

An elaborate formula which can be used in case of equal ranks is:

$$\rho = 1 - \frac{6}{n^3 - n}\left[\Sigma D^2 + \frac{1}{12}\Sigma(m^3 - m)\right].$$

Here, $\frac{1}{12}\Sigma(m^3 - m)$ is to be added to ΣD^2 for each group of equal ranks, m being the number of equal ranks each time.

For the given data, you have:

For series x, the number of equal ranks $m = 2$

For series y, also, $m = 2$; so that,

$$\rho = 1 - \frac{6}{5^3 - 5}\left[21 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]$$

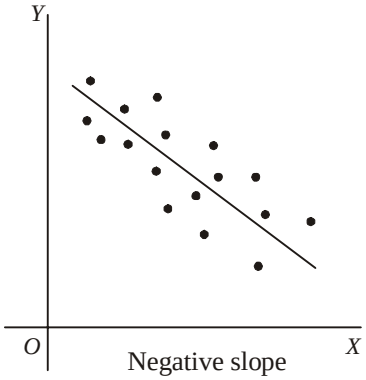
$$= 1 - \frac{6}{120}\left[21 + \frac{6}{12} + \frac{6}{12}\right]$$

$$= 1 - \frac{6 \times 22}{120} = -0.1$$

Example 3.7: Show by means of diagrams various cases of scatter expressing correlation between x, y.

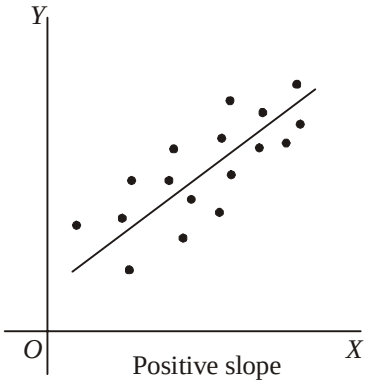
Solution:

(a)



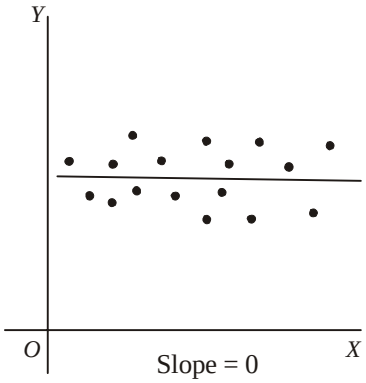
Negative slope
Inverse linear relationship
High scatter r low, negative

(b)



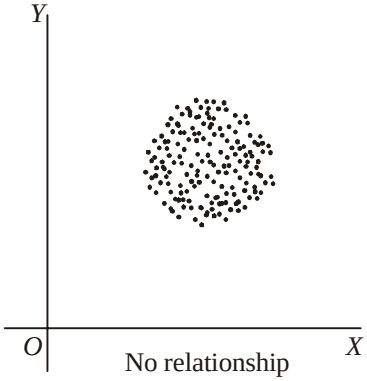
Positive slope
Direct linear relationship
High scatter r low, positive

(c)



Slope = 0
$r = 0$

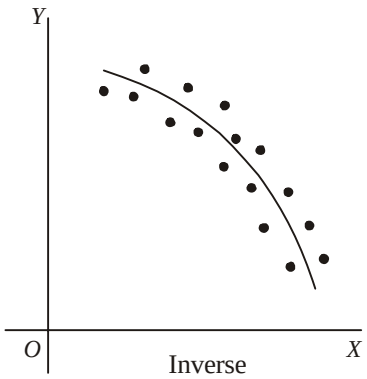
(d)



No relationship
$r \simeq 0$

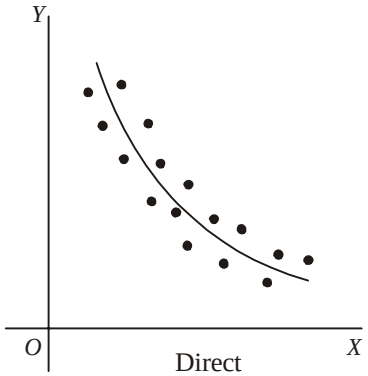
NOTES

(e)



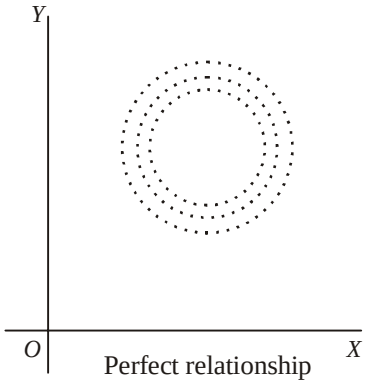
Inverse
curvilinear relationship

(f)



Direct
curvilinear relationship

(g)



Perfect relationship
But, $r = 0$ because of
non-linear relation

Correlation analysis helps us in determining the degree to which two or more variables are related to each other. When there are only two variables, you can determine the degree to which one variable is linearly related to the other.

CHECK YOUR PROGRESS

1. What do you understand by correlation analysis?
2. When is a correlation positive and when is it negative?
3. What do you understand by the coefficient of correlation?
4. How do we calculate the coefficient of correlation?
5. What do you understand by the coefficient of non-determination and coefficient of alienation?
6. State the precautionary measures to be taken in using regression and correlation analyses.

3.3 REGRESSION ANALYSIS

The term 'regression' was first used in 1877 by Sir Francis Galton who made a study that showed that the height of children born to tall parents would tend to move back or 'regress' toward the mean height of the population. He

designated the word regression as the name of the process of predicting one variable from the another variable. He coined the term multiple regression to describe the process by which several variables are used to predict another. Thus, when there is a well established relationship between variables, it is possible to make use of this relationship in making estimates and to forecast the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s). A banker, for example, could predict deposits on the basis of per capita income in the trading area of the bank. A marketing manager may plan his advertising expenditures on the basis of the expected effect on total sales revenue of a change in the level of advertising expenditure. Similarly, a hospital superintendent could project his need for beds on the basis of total population. Such predictions may be made by using regression analysis. An investigator may employ regression analysis to test his theory having the cause and effect relationship. All this explains that regression analysis is an extremely useful tool, especially in business and industry problems involving predictions.

Assumptions in regression analysis

While using of the regression techniques for making predictions, it is always assumed:

- There is an actual relationship between the dependent and independent variables.
- The values of the dependent variable are random but the values of the independent variable are fixed quantities without error and are chosen by the experimentor.
- There is clear indication of direction of the relationship. This means that dependent variable is a function of independent variable. (For example, when you say that advertising has an effect on sales, then you are saying that sales has an effect on advertising).
- The conditions (that existed when the relationship between the dependent and independent variables was estimated by the regression) are the same when the regression model is being used. In other words, it simply means that the relationship has not changed since the regression equation was computed.
- The analysis can be used to predict values within the range (and not for values outside the range) for which it is valid.

3.3.1 Simple Linear Regression Model

In case of simple linear regression analysis, a single variable is used to predict another variable on the assumption of linear relationship (i.e. relationship of the type defined by $Y = a + bX$) between the given variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.

NOTES

Simple linear regression model[4] (or the regression line) is stated as

$$Y_i = a + bX_i + e_i$$

where Y_i is the dependent variable

X_i is the independent variable

e_i is unpredictable random element (usually called as residual or error term)

(a) *a* represents the *Y*-intercept, i.e. the intercept specifies the value of the dependent variable when the independent variable has a value of zero. (But this term has practical meaning only if a zero value for the independent variable is possible).

(b) *b* is a constant indicating the slope of the regression line. The slope of the line indicates the amount of change in the value of the dependent variable for a unit change in the independent variable.

If the two constants (namely *a* and *b*) are known, the accuracy of your prediction of *Y* (denoted by $\hat{Y}$ and read as *Y*-hat) depends on the magnitude of the values of e_i. If in the model, all the e_i tend to have very very large values, then the estimates will not be very good, but if these values are relatively small, then the predicted values ($\hat{Y}$) will tend to be close to the true values (Y_i).

3.3.2 Estimating the Intercept and Slope of the Regression Model (or Estimating the Regression Equation)

The two constants or the parameters, namely '*a*' and '*b*' in the regression model for the entire population or universe are generally unknown and as such are estimated from sample information. The following are the two methods used for estimation:

(*i*) Scatter diagram method

(*ii*) Least squares method

(i) Scatter diagram method

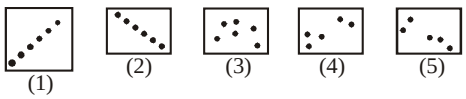
This method makes use of the Scatter diagram also known as Dot diagram. *Scatter diagram*[5] is a diagram representing two series with the known variable,

---

[4] Usually, the estimate of *Y* denoted by $\hat{Y}$ is written as

$$\hat{Y} = a + bX_i$$

on the assumption that the random disturbance to the system averages out or has an expected value of zero (i.e. $e = 0$) for any single observation. This regression model is known as the regression line of *Y* on *X* from which the value of *Y* can be estimated for the given value of *X*.

[5] 

Five possible forms which a scatter diagram may assume have been depicted in the five diagrams. The *first* diagram is indicative of a perfect positive relationship; the *second*

i.e. independent variable plotted on the *X*-axis and the variable to be estimated, i.e. dependent variable to be plotted on the *Y*-axis on a graph paper (refer to Figure 3.2 and Table 3.2) to get the following information.

Table 3.2 *Income and Consumption Data*

Income *X* *(Hundreds of Rupees)*	*Consumption Expenditure* *Y* *(Hundreds of Rupees)*
41	44
65	60
50	39
57	51
96	80
94	68
110	84
30	34
79	55
65	48

The scatter diagram by itself is not sufficient for predicting values of the dependent variable. Some formal expression of the relationship between the two variables is necessary for predictive purposes. For the purpose, one may simply take a ruler and draw a straight line through the points in the scatter diagram and this way can determine the intercept and the slope of the said line and then the line can be defined as $\hat{Y} = a + bX_i$ with the help of which we can predict *Y* for a given value of *X*. But there are shortcomings in this approach. For example, if five different persons draw such a straight line in the same scatter diagram, it is possible that there may be five different estimates of *a* and *b*, specially when the dots are more dispersed in the diagram. Hence, the estimates cannot be worked out only through this approach. A more systematic and statistical method is required to estimate the constants of the predictive equation. The least squares method is used to draw the best fit line.

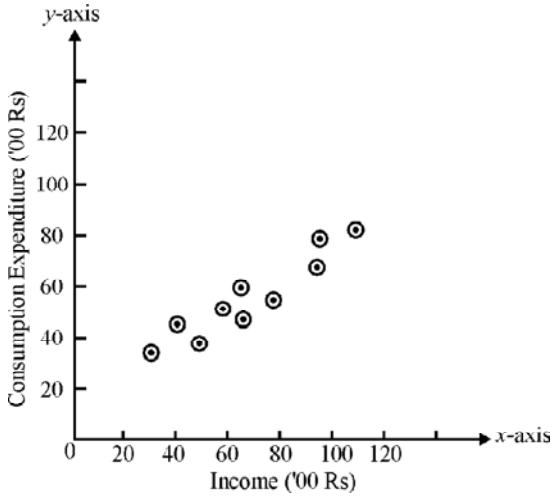


Fig. 3.2 *Scatter Diagram*

shows a perfect negative relationship; the *third* shows no relationship; the *fourth* shows a positive relationship; and the *fifth* shows a negative relationship between the two variables under consideration.

(ii) Least squares method

The least squares method of fitting a line (the line of best fit or the regression line) through the scatter diagram is a method which minimizes the sum of the squared vertical deviations from the fitted line. In other words, the line to be fitted will pass through the points of the scatter diagram in such a way that the sum of the squares of the vertical deviations of these points from the line will be a minimum.

The meaning of the least squares criterion can easily be understood through reference to Figure 3.3, where Figure 3.2 has been reproduced along with a line which represents the least squares line fit to the data.

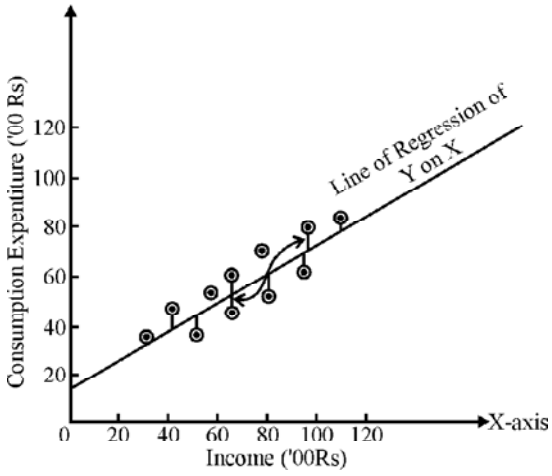


Fig. 3.3 *Scatter Diagram, Regression Line and Short Vertical Lines Representing 'e'*

In Figure 3.3, the vertical deviations of the individual points from the line are shown as the short vertical lines joining the points to the least squares line. These deviations will be denoted by the symbol '*e*'. The value of '*e*' varies from one point to another. In some cases it is positive, while in others it is negative. If the line drawn happens to be least squares line, then the values of Σe_i is the least possible. It is because of this feature the method is known as the least squares method.

Why you should insist on minimizing the sum of squared deviations is a question that needs explanation. If we denote the deviations from the actual value Y to the estimated value $\hat{Y}$ as $(Y - \hat{Y})$ or e_i, it is logical that we want the $\Sigma(Y - \hat{Y})$ or $\sum_{i=1}^{n} e_i$, to be as small as possible. However, mere examining $\Sigma(Y - \hat{Y})$ or $\sum_{i=1}^{n} e_i$, is inappropriate, since any e_i can be positive or negative. Large positive values and large negative values could cancel one another. But large values of e_i regardless of their sign, indicate a poor prediction. Even if we

ignore the signs while working out $\sum_{i=1}^{n} |e_i|$ the difficulties may continue. Hence, the standard procedure is to eliminate the effect of signs by squaring each observation. Squaring each term accomplishes two purposes, namely (*a*) it magnifies (or penalizes) the larger errors and (*b*) it cancels the effect of the positive and negative values (since a negative error when squared becomes positive). The choice of minimizing the squared sum of errors rather than the sum of the absolute values implies that there are many small errors rather than a few large errors. Hence, in obtaining the regression line we follow the approach that the sum of the squared deviations be minimum and on this basis work out the values of its constants, namely '*a*' and '*b*' also known as the intercept and the slope of the line. This is done with the help of the following two normal equations:[6]

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

In the above two equations, '*a*' and '*b*' are unknowns and all other values, viz. ΣX, ΣY, ΣX^2, ΣXY are the sum of the products and cross products to be calculated from the sample data, and '*n*' means the number of observations in the sample.

The following examples explain the least squares method.

Example 3.8: Fit a regression line $\hat{Y} = a + bX_i$ by the method of least squares to the given sample information.

Observations	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Income (*X*) ('00 Rs)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (*Y*) ('00 Rs)	44	60	39	51	80	68	84	34	55	48

Solution: You are to fit a regression line $\hat{Y} = a + bX_i$ to the given data by the method of least squares. Accordingly, work out the '*a*' and '*b*' values with the help of the normal equations as stated above and also for the purpose work out ΣX, ΣY, ΣXY, ΣX^2 values from the given sample information table on summations for regression equation.

---

[6] If you proceed centring each variable, i.e. setting its origin at its mean, then the two equations will be as follows:

$$\Sigma Y = na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

But since ΣY and ΣX will be zero, the first equation and the first term of the second equation will disappear and you shall simply have the following equations:

$$\Sigma XY = b\Sigma X^2$$
$$b = \Sigma XY/\Sigma X^2$$

The value of '*a*' can then be worked out as

$$a = \bar{Y} - b\bar{X}$$

NOTES

Summations for Regression Equation

Observations	*Income X* (*'00 Rs*)	*Consumption Expenditure Y* (*'00 Rs*)	*XY*	X^2	Y^2
1	41	44	1804	1681	1936
2	65	60	3900	4225	1600
3	50	39	1950	2500	1521
4	57	51	2907	3249	2601
5	96	80	7680	9216	6400
6	94	68	6392	8836	4624
7	110	84	9240	12100	7056
8	30	34	1020	900	1156
9	79	55	4345	6241	3025
10	65	48	3120	4225	2304
$n = 10$	$\sum X = 687$	$\sum Y = 563$	$\sum XY = 42358$	$\sum X^2 = 53173$	$\sum Y^2 = 34223$

Putting the values in the required normal equations, you have:

$$563 = 10a + 687b$$

$$42358 = 687a + 53173b$$

Solving these two equations for a and b, you obtain:

$$a = 14.000 \quad \text{and} \quad b = 0.616$$

Hence, the equation for the required regression line is:

$$\hat{Y} = a + bX_i$$

or

$$\hat{Y} = 14.000 + 0.616X_i$$

This equation is known as the regression equation of Y on X from which Y values can be estimated for given values of X variable.[7]

3.3.3 Checking the Accuracy of Equation

After finding the regression line as stated above, one can check its accuracy also. The method to be used for the purpose follows from the mathematical property of a line fitted by the method of least squares, namely the individual positive and negative errors must sum to zero. In other words, using the estimating equation one must find out whether the term $\sum\left(Y - \hat{Y}\right)$ is zero and if this is so, then one can reasonably be sure that he has not committed any mistake in determining the estimating equation.

[7] It should be pointed out that the equation used to estimate the Y variable values from values of X should not be used to estimate the values of variable X from the given values of variable Y. Another regression equation (known as the regression equation of X on Y of the type $X = a + bY$) that reverses the two values should be used if it is desired to estimate X from the value of Y.

Problem of prediction

When we talk about prediction or estimation, we usually imply that if the relationship $Y_i = a + bX_i + e_i$ exists then the regression equation, $\hat{Y} = a + bX_i$ provides a basis for making estimates of the value for Y which will be associated with particular values of X. In Example 3.8, we worked out the regression equation for the income and consumption data as

$$\hat{Y} = 14.000 + 0.616X_i$$

On the basis of this equation, we can make a *point estimate* of Y for any given value of X. Suppose we wish to estimate the consumption expenditure of individuals with income of Rs 10,000. We substitute $X = 100$ for the same in our equation and get an estimate of consumption expenditure as follows:

$$\hat{Y} = 14.000 + 0.616(100) = 75.60$$

Thus, the regression relationship indicates that individuals with Rs 10,000 of income may be expected to spend approximately Rs 7560 on consumption. But this is only an expected or an estimated value and it is possible that actual consumption expenditure of same individual with that income may deviate from this amount and if so, then our estimate will be an error, the likelihood of which will be high if the estimate is applied to any one individual. The *interval estimate* method is considered better and it states an interval in which the expected consumption expenditure may fall. Remember that the wider the interval, the greater the level of confidence we can have, but the width of the interval (or what is technically known as the precision of the estimate) is associated with a specified level of confidence and is dependent on the variability (consumption expenditure in our case) found in the sample. This variability is measured by the standard deviation of the error term, '*e*', and is popularly known as the standard error of the estimate.

Standard error of the estimate

Standard error of estimate is a measure developed by the statisticians for measuring the reliability of the estimating equation. Like the standard deviation, the Standard Error (SE) of $\hat{Y}$ measures the variability or scatter of the observed values of Y around the regression line. Standard Error of Estimate (SE of $\hat{Y}$) is worked out as under:

$$\text{SE of } \hat{Y} \text{ (or } S_e) = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{\sum e^2}{n-2}}$$

where

SE of $\hat{Y}$ (or S_e) = Standard error of the estimate

Y = Observed value of Y

$\hat{Y}$ = Estimated value of Y

e = The error term = $(Y - \hat{Y})$

n = Number of observations in the sample

NOTES

Note: In the preceding formula, $n-2$ is used instead of n because of the fact that two degrees of freedom are lost in basing the estimate on the variability of the sample observations about the line with two constants, namely '*a*' and '*b*', whose position is determined by those same sample observations.

The square of the S_e also known as the variance of the error term is the basic measure of reliability. The larger the variance, the more significant are the magnitudes of the *e*'s and the less reliable is the regression analysis in predicting the data.

Interpreting the standard error of estimate and finding the confidence limits for the estimate in large and small samples

The larger the ΣE of estimate (SE_e), the greater happens to be the dispersion, or scattering, of given observations around the regression line. But if the ΣE of estimate happens to be zero then the estimating equation is a 'perfect' estimator (i.e. cent per cent correct estimator) of the dependent variable.

In case of large samples, i.e. where $n > 30$ in a sample, it is assumed that the observed points are normally distributed around the regression line and we may find

68% of all points within $\hat{Y} \pm 1\ SE_e$ limits

95.5% of all points within $\hat{Y} \pm 2\ SE_e$ limits

99.7% of all points within $\hat{Y} \pm 3\ SE_e$ limits

This can be stated as follows:

(a) The observed values of *Y* are normally distributed around each estimated value of $\hat{Y}$.

(b) The variance of the distributions around each possible value of $\hat{Y}$ is the same.

In case of small samples, i.e. where $n \leq 30$ in a sample the '*t*' distribution is used for finding the two limits more appropriately.

This is done as follows:

$$\text{Upper limit} = \hat{Y} + \text{'}t\text{'}\ (SE_e)$$

$$\text{Lower limit} = \hat{Y} - \text{'}t\text{'}\ (SE_e)$$

where

$\hat{Y}$ = The estimated value of *Y* for a given value of *X*.

SE_e = The standard error of estimate.

'*t*' = Table value of '*t*' for given degrees of freedom for a specified confidence level.

NOTES

3.3.4 Some Other Details Concerning Simple Regression

Sometimes the estimating equation of *Y*, also known as the regression equation of *Y* on *X*, is written as follows:

$$\left(\hat{Y} - \overline{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right)$$

or

$$\hat{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right) + \overline{Y}$$

where

r = Coefficient of simple correlation between X and Y

σ_Y = Standard deviation of Y

σ_X = Standard deviation of X

$\overline{X}$ = Mean of X

$\overline{Y}$ = Mean of Y

$\hat{Y}$ = Value of Y to be estimated

X_i = Any given value of X for which Y is to be estimated.

This is based on the formula we have used, i.e. $\hat{Y} = a + bX_i$. The coefficient of X_i is defined as

$$\text{Coefficient of } X_i = b = r\frac{\sigma_Y}{\sigma_X}$$

(Also known as regression coefficient of *Y* on *X* or slope of the regression line of *Y* on *X*) or b_{YX}.

$$= \frac{\sum XY - n\overline{X}\,\overline{Y} \times \sqrt{\sum Y^2 - n\overline{Y}^2}}{\sqrt{\sum Y^2 - n\overline{Y}^2}\sqrt{\sum X^2 - n\overline{X}^2}\sqrt{\sum X^2 - n\overline{X}^2}}$$

$$= \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

and

$$a = -r\frac{\sigma_Y}{\sigma_X}\overline{X} + \overline{Y}$$

$$= \overline{Y} - b\overline{X} \qquad \left(\text{since } b = r\frac{\sigma_Y}{\sigma_X}\right)$$

Similarly, the estimating equation of *X* also known as the regression equation of *X* on *Y* can be stated as follows:

NOTES

$$\left(\hat{X} - \overline{X}\right) = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right)$$

or

$$\hat{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right) + \overline{X}$$

and

$$\text{Regression coefficient of } X \text{ on } Y \text{ (or } b_{XY}) = r\frac{\sigma_X}{\sigma_Y} = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum Y^2 - n\overline{Y}^2}$$

If we are given the two regression equations as stated above along with the values of '*a*' and '*b*' constants to solve the same for finding the value of *X* and *Y*, then the values of *X* and *Y* so obtained are the mean value of *X* (i.e. $\overline{X}$) and the mean value of *Y* (i.e. $\overline{Y}$).

If we are given the two regression coefficients (namely b_{XY} and b_{YX}) then we can work out the value of coefficient of correlation by just taking the square root of the product of the regression coefficients as shown in the following:

$$r = \sqrt{b_{YX}.b_{XY}}$$

$$= \sqrt{r\frac{\sigma_Y}{\sigma_X}.r\frac{\sigma_X}{\sigma_Y}}$$

$$= \sqrt{r.r} = r$$

The (±) sign of *r* will be determined on the basis of the sign of the regression coefficients given. If regression coefficients have minus sign then *r* will be taken with minus (–) sign and if regression coefficients have plus sign then *r* will be taken with plus (+) sign. (Remember that both regression coefficients will necessarily have the same sign whether it is minus or plus for their sign is governed by the sign of coefficient of correlation.)

Example 3.9: Given is the following information:

	$\overline{X}$	$\overline{Y}$
Mean	39.5	47.5
Standard Deviation	10.8	17.8

Simple correlation coefficient between *X* and *Y* is = + 0.42

Find the estimating equation of *Y* and *X*.

Solution: Estimating equation of *Y* can be worked out as follows:

$\because$

$$\left(\hat{Y} - \overline{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right)$$

or

$$\hat{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right) + \overline{Y}$$

$$= 0.42\frac{17.8}{10.8}\left(X_i - 39.5\right) + 47.5$$

NOTES

$$= 0.69X_i - 27.25 + 47.5$$

$$= 0.69X_i + 20.25$$

Similarly, the estimating equation of X can be worked out as under:

$$\because \quad \left(\hat{X} - \overline{X}\right) = r\frac{\sigma_X}{\sigma_Y}\left(Y_i - \overline{Y}\right)$$

or $$\hat{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y_i - \overline{Y}\right) + \overline{X}$$

or $$= 0.42\frac{10.8}{17.8}\left(Y_i - 47.5\right) + 39.5$$

$$= 0.26Y_i - 12.25 + 39.5$$

$$= 0.26Y_i + 27.25$$

Example 3.10: Given is the following data:

Variance of $X = 9$

Regression equations:

$$4X - 5Y + 33 = 0$$

$$20X - 9Y - 107 = 0$$

Find: (a) Mean values of X and Y.

(b) Coefficient of Correlation between X and Y.

(c) Standard deviation of Y.

Solution:

(a) For finding the mean values of X and Y we solve the two given regression equations for the values of X and Y as follows:

$$4X - 5Y + 33 = 0 \qquad \text{(i)}$$

$$20X - 9Y - 107 = 0 \qquad \text{(ii)}$$

If we multiply equation (1) by 5 we have the following equations:

$$20X - 25Y = -165 \qquad \text{(iii)}$$

$$20X - 9Y = 107 \qquad \text{from (ii)}$$

$$- \quad + \quad -$$

$$-16Y = -272 \qquad \text{Subtracting equation (ii) from (iii)}$$

or $$Y = 17$$

Putting this value of Y in equation (i) you have:

$$4X = -33 + 5(17)$$

or $$X = \frac{-33 + 85}{4} = \frac{52}{4} = 13$$

Hence, $$\overline{X} = 13 \quad \text{and} \quad \overline{Y} = 17$$

NOTES

(b) For finding the coefficient of correlation, first of all you presume one of the two given regression equations as the estimating equation of X. Let equation $4X - 5Y + 33 = 0$ be the estimating equation of X, then you have:

$$\hat{X} = \frac{5Y_i}{4} - \frac{33}{4}$$

and

From this, you can write $b_{XY} = \frac{5}{4}$

The other given equation is then taken as the estimating equation of Y and can be written as follows:

$$\hat{Y} = \frac{20X_i}{9} - \frac{107}{9}$$

and from this, you can write $b_{YX} = \frac{20}{9}$

If the preceding equations are correct, then r must be equal to

$$r = \sqrt{5/4 \times 20/9} = \sqrt{25/9} = 5/3 = 1.6$$

Which is an impossible equation, since r can in no case be greater than 1. Hence, you change our supposition about the estimating equations and by reversing it, you rewrite the estimating equations as under:

$$\hat{X} = \frac{9Y_i}{20} + \frac{107}{20}$$

and

$$\hat{Y} = \frac{4X_i}{5} + \frac{33}{5}$$

Hence,

$$r = \sqrt{9/20 \times 4/5}$$

$$= \sqrt{9/25}$$

$$= 3/5 = 0.6$$

Since, regression coefficients have plus signs, you take $r = + 0.6$

(c) Standard deviation of Y can be calculated as follows:

$\because$ Variance of $X = 9$ $\quad\quad$ $\therefore$ Standard deviation of $X = 3$

$\because$ $$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{4}{5} = 0.6\frac{\sigma_Y}{3} = 0.2\sigma_Y$$

$$\text{Hence } \sigma_Y = 4$$

Alternatively we can work it out as under:

$\because$ $$b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{9}{20} = 0.6\frac{\sigma_Y}{3} = \frac{1.8}{\sigma_Y}$$

$$\text{Hence } \sigma_Y = 4$$

NOTES

3.3.5 Regression of Two Lines

The relationship between Y and X is not perfect. The average relationship of Y and X in which X is the independent and Y the dependent variable is not the same as the average relationship of X and Y in which Y is the independent and X the dependent variable.

The two regression lines, which best describe these two average relationships, are given by the regression equations (Figure 3.4):

$$Y = a + bX$$

$$X = a' + b'Y$$

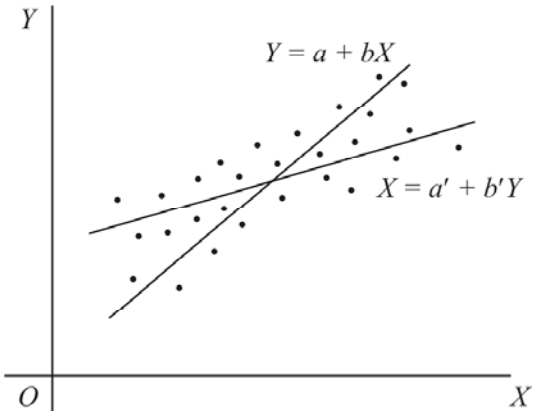


Fig. 3.4 *Scatter Diagram with Lines of Best Fit*

a, b are obtained by minimizing $\Sigma(Y - Y')^2$

a', b' are obtained by minimizing $\Sigma(X - X')^2$

$$b = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

$$b' = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma Y^2 - n\bar{Y}^2}$$

$$a = Y - b\bar{X}$$

$$a' = \bar{X} - b'\bar{Y}$$

The signs of b and b' indicate whether the slopes of the lines of best fit are positive or negative. It may be recalled that regression coefficient measures the average change in the dependent variable corresponding to a unit change in the independent variable.

b and b' must have the same sign, both + or both –. Also, r has the same sign as b, b'.

A positive value of the regression coefficient indicates that the relation between X and Y is direct.

NOTES

A negative value shows an inverse relationship between X and Y, i.e. high and low values are paired together.

Since $b = r\frac{s_y}{s_x} = b_{yx}$ $\qquad$ $b' = r\frac{s_x}{s_y} = b_{xy}$

We have $bb' = r^2$ or $b_{yx}b_{xy} = r^2$

That r, b, b' have the same sign gives us an alternative definition of r. It is the square root of the product of b and b' and has the same sign as b or b'.

$$r = \sqrt{bb'} \text{ or } \sqrt{b_{yx}b_{xy}}$$

The product bb' can never exceed 1 because r cannot numerically exceed 1. This result can be used to distinguish between two regression lines for the same data.

For example, you have $b = -0.48$, $b' = -1.66$

$$\therefore \qquad r = \sqrt{(-0.48)(-1.66)} = -0.89$$

If the regression line of Y on X, i.e. $Y = a + bX$ exists it does not imply that regression of X on Y necessarily exists.

If X is time and Y sales, the regression of sales on time is expressed by $Y = a + bX$. But there is no question of regression (or dependence) of time on sales.

Can the two regression lines coincide?

The two regression lines are identical if and only if all the points in the scatter diagram lie on one straight line, i.e. if the correlation is perfect, i.e. $r = 1$.

What is the point of intersection of the two regression lines?

$(\bar{X}, \bar{Y})$ is the only point common to both and hence, the point of intersection. If you solve the two regression equations simultaneously, you get $\bar{X}$, $\bar{Y}$.

Example 3.11: For the following data showing index numbers of prices and production for five years, find the two regression lines and show that $bb' = r^2$.

Year	*Index Numbers of*	
	Production	*Prices*
1961	100	107
1962	101	123
1963	106	133
1964	99	109
1965	97	128

Solution: Estimate the index number of prices when it is known that the index number of production is 110. Predict the index number of production when that of prices is known to be 120. Use X for production, Y for prices.

Subtract 100 from each value of X and 120 from each value of Y.

Note: *r* does not change by the change of scale and origin, i.e. by subtraction and division. *b* changes by division (or multiplication) of observations by any number. But *b* does not change by subtraction as done in this exercise. In simple regression problems, such subtraction may be avoided.

Thus, $u = X - 100$, $v = Y - 120$

u	v	u^2	v^2	uv
0	–13	0	169	0
1	3	1	9	3
6	13	36	169	78
–1	–11	1	121	11
–3	9	9	81	–27
$\Sigma u = 3$	$\Sigma v = 1$	$\Sigma u^2 = 47$	$\Sigma v^2 = 549$	$\Sigma uv = 65$

$$\bar{u} = \frac{\Sigma u}{n} = \frac{3}{5} = 0.6,\ \bar{v} = \frac{\Sigma v}{n} = \frac{1}{5} = 0.2$$

(*i*)
$$r = \frac{\Sigma uv - n\bar{u}\bar{v}}{\sqrt{\Sigma u^2 - n\bar{u}^2}\sqrt{\Sigma v^2 - n\bar{v}^2}}$$

$$= \frac{65 - 5 \times 0.6 \times 0.2}{\sqrt{147 - 5 \times 0.6 \times 0.6}\sqrt{549 - 5 \times 0.2 \times 0.2}}$$

$$= \frac{64.4}{\sqrt{45.2}\sqrt{548.8}}$$

$$= \frac{64.4}{6.3 \times 23.4} = 0.41$$

(*ii*) For the regression of *Y* on *X*, no additional work is necessary. *b* does not change by change of origin only. It changes by change of scale:

$$b = \frac{\Sigma uv - n\bar{u}\bar{v}}{\Sigma u^2 - n\bar{u}^2} = \frac{64.4}{45.2} = 1.42$$

Now,
$$\bar{X} = \frac{\Sigma X}{n} = \frac{503}{5} = 100.6 \text{ and } \bar{Y} = \frac{\Sigma Y}{n} = \frac{601}{5} = 120.2$$

$\therefore$
$$a = \bar{Y} - b\bar{X} = 120.2 - 1.42 \times 100.6 = -14.25$$

The regression of *Y* on *X* is given by

$Y = -14.25 + 1.42\ X$

(*iii*) To find the regression of *X* on *Y*

$$b' = \frac{\Sigma uv - n\bar{u}\bar{v}}{\Sigma v^2 - n\bar{v}^2} = \frac{64.4}{548.8} = 0.12$$

$$a' = \bar{X} - b'\ \bar{Y} = 100.6 - 0.12 \times 120.2 = 68.58$$

The regression of *X* on *Y* is given by:

$$X = 68.58 + 0.12\ Y$$

(*iv*)
$$\sqrt{bb'} = \sqrt{1.42 \times 0.12} = 0.41 = r \qquad \because\ bb' = r^2$$

NOTES

(*v*) To predict Y from X, substitute $X = 110$ in the regression of Y on X.

Predicted $Y = -14.25 + 1.42 \times 110 = 141.95$

To predict X from Y substitute $Y = 120$ in the regression of X on Y.

Predicted $X = 68.58 + 0.12 \times 120 = 82.98$

Example 3.12: A firm doubles the number of its employees and profit increases significantly. Does it imply that profit depends on the number of employees?

Solution: It is likely that the increase in the employee number has come along with increase in capital, efficiency or other development. The increase in employee number need not be the basic cause of increase in profit.

If X and Y have nothing to do with each other logically but the observations on X and Y happen to move according to a pattern, the resulting regression equation, even if it is well fitted and has significant coefficients, is spurious and meaningless.

Example 3.13: In a linear regression analysis of 60 observations, the two lines of regression are:

$$1000\ Y = 768X - 3608 \text{ and } 5X = 6Y + 24$$

What is the coefficient of correlation in the data? Also show that the ratio of the coefficient of variation of X to that of Y is $\frac{5}{24}$.

Solution: From the data you have

$$b = \frac{768}{1000},\ b' = \frac{6}{5}$$

Coefficient of correlation $r = \sqrt{bb'} = \sqrt{0.922} = 0.96$

If we solve the two equations, we get $\bar{X} = 6, \bar{Y} = 1$.

$$\text{Since } b = r\frac{s_y}{s_x} \quad \therefore \frac{s_x}{s_y} = \frac{r}{b} = \frac{0.96}{0.768} = 1.25$$

Coefficient of variation of X is $\frac{s_x}{\bar{X}} \times 100$

Coefficient of variation of Y is $\frac{s_y}{\bar{Y}} \times 100$

Their ratio is $\frac{s_x / \bar{X}}{s_y / \bar{Y}} = \frac{s_x}{s_y}\frac{\bar{Y}}{\bar{X}} = 1.25 \times \frac{1}{6} = \frac{5}{24}$

Example 3.14: What if $bb' > 1$?

Solution: Since $r^2 \not> 1$, $\therefore$ If $bb' > 1$, interchange dependent and independent variables in the two regression lines.

Example 3.15: The equations of two regression lines obtained in a correlation analysis of 60 observations are $5x = 6y + 24$ and $1000y = 768x - 3608$. What is the correlation coefficient and what is its probable error? Also show that the ratio of the coefficient of variance of x to that of y is $\frac{5}{24}$. What is the ratio of variance of x and y?

Solution: The equations of the regression lines are given as follows:

$5x = 6y + 24$ and $1000y = 768x - 3608$

$$\therefore \quad b_{xy} = r.\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \qquad \text{(i)}$$

$$\text{and} \quad b_{yx} = r.\frac{\sigma_y}{\sigma_x} = \frac{768}{1000} \qquad \text{(ii)}$$

Multiplying equations (i) and (ii), you get

$$b_{xy} \times b_{yx} = r^2 = \frac{6}{5} \times \frac{768}{1000} \Rightarrow r = \pm 0.96$$

Since both b_{xy} and b_{yx} are positive, the correlation coefficient r is also positive and hence, $r = + 0.96$.

Probable error of r,

$$PE_r = 0.6745\left(\frac{1-r^2}{\sqrt{N}}\right)$$

$$PE_r = 0.6745\left(\frac{1-0.96^2}{\sqrt{60}}\right)$$

Each regression line passes through $(\bar{x}, \bar{y})$. So, from the given equations of these lines you have

$$5\bar{x} = 6\bar{y} + 24$$

$$\text{and} \quad 1000\bar{y} = 768\bar{x} - 3608$$

Solving these, you get

$$\bar{x} = 6 \text{ and } \bar{y} = 1 \qquad \text{(iii)}$$

From equation (i), you have $r.\frac{\sigma_x}{\sigma_y} = \frac{6}{5}$, where $r = 0.96$.

$$\text{or} \quad \frac{\sigma_x}{\sigma_y} = \frac{6}{5} \times \frac{1}{0.96} = \frac{5}{4} \qquad \text{(iv)}$$

And the ratio of the coefficients of variance of x to that of y

$$\frac{(\sigma_x / \bar{x})}{(\sigma_y / \bar{y})} = \left(\frac{\bar{y}}{\bar{x}}\right) \times \left(\frac{\sigma_x}{\sigma_y}\right) = \left(\frac{1}{6}\right) \times \left(\frac{5}{4}\right)$$

[from equations (3) and (4)]

$$= \frac{5}{24}$$

NOTES

Example 3.16: Two lines of regression are $x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$ and variance of x is 12. Calculate the values of $\bar{x}, \bar{y}, \sigma_y^2$ and r.

Solution: Since each regression line passes through $(\bar{x}, \bar{y})$, so from the given equations, you have $2\bar{y} = -\bar{x} + 5$

and $2\bar{x} = -3\bar{y} + 8$.

Solving these, you get $\bar{x} = 1$, $\bar{y} = 2$.

Assuming the lines of regression of y on x and x on y as

$$2y = -x + 5 \text{ and } 2x = -3y + 8 \quad \text{(i)}$$

From equation (i) you have

$$b_{xy} = r\frac{\sigma_y}{\sigma_x} = -\frac{1}{2} \quad \text{(ii)}$$

and $$b_{xy} = r\frac{\sigma_x}{\sigma_y} = -\frac{3}{2}$$

Multiplying equations (i) and (ii), you get

$$b_{xy} \times b_{yx} = \left(-\frac{3}{2}\right) \times \left(-\frac{1}{2}\right) = r^2$$

$$\Rightarrow \quad r^2 = \frac{3}{4} \Rightarrow r = \pm\frac{\sqrt{3}}{2} = \pm\, 0.866$$

Since b_{xy} and b_{yx} are negative, the correlation coefficient r is negative.

Thus, $r = -0.866$

Now, $\sigma_x^2 = 12$ (given)

From equation (ii), you have $\left(\frac{r\,\sigma_x}{\sigma_y}\right)^2 = \left(-\frac{1}{2}\right)^2$

or $$\frac{r^2\,\sigma_x^2}{\sigma_y^2} = \frac{1}{4}$$

or $4(-0.866)2 \times 12 = \sigma_y^2$

$$\Rightarrow \quad \sigma_y^2 = 35.998$$

Note: If we assume the lines of regression of y on x and x on y as. $x = -2y + 5$ and $3y = -2x + 8$, then you shall get

$r^2 = b_{xy} \times b_{xy} = \left(-\frac{2}{3}\right)(-2) = \frac{4}{3} > 1$, which is inadmissible.

CHECK YOUR PROGRESS

7. What is regression analysis?
8. What are the assumptions involved in using regression analysis for making predictions?
9. What is a simple linear regression model?
10. Define the two constants involved in regression.
11. Name the methods to calculate the constants in regression models.
12. What is the scatter diagram method?
13. What is the least squares method?
14. What is the standard error of the estimate?

3.4 RELATIONSHIP BETWEEN CORRELATION AND REGRESSION COEFFICIENTS

Regression analysis is the statistical technique used to evaluate the phenomena and to predict the future events. It defines the relationship or the correlation between one dependent variable and several independent variables. Regression is analysed for partial correlation as well as multiple correlation. Thus, in regression analysis, a coefficient of correlation *r* between any two random variables *X* (predictor variable) and *Y* (criterion variable) is termed as the quantitative index of association between these two variables. When the coefficient of correlation *r* is squared, it becomes a coefficient of determination r^2, which describes the amount of variance between the variables *X* and *Y*. Hence, in a multiple regression analysis the predictor variables $X_1, X_2, ..., X_n$ describe the variability of criterion variable *Y*. The coefficient of multiple determination is R^2 and the coefficient of multiple correlation is *R*.

Hence, the multiple regression analysis is a statistical tool that helps researchers to evaluate the effect of different factors on the consequences occurring at the same time. It analyses the relationship between several independent or predictor variables and a dependent variable. In research, regression analysis is used to investigate a particular set of predictors and to show differences in the consequences that occur. Generally, regression is used to determine the effect of the specific factors along with the other factors that influence these consequences. Researchers use algebraic methods to analyse the result by making a group of factors associated with a particular phenomenon as a constant. According to the dictionary meaning, multiple regression is a statistical technique that predicts values of one variable on the basis of two or more other variables.

Multiple regression and statistics: The term multiple regression was first given by Pearson. Regression is of two types, simple and multiple, and both the regression techniques are related to the analysis of variance (ANOVA). Multiple

NOTES

regression is the simplest method in comparison to other multivariate statistical techniques.

Multiple regression and mathematics: The multiple regression technique is used in mathematics to formulate simple regression equations, and to evaluate the best fitting curve for straight line along the dots on an *x-y* plot or a scatter diagram.

Rules for multiple regression analysis

The following rules should be followed in the multiple regression analysis.

- In multiple regression analysis, one dependent variable and many independent variables are considered, while in simple regression only one independent variable is considered. In factor analysis and other multivariate techniques, the dependent variables are evaluated.
- In multiple regression analysis, you evaluate and correlate the independent variables.
- In multiple regression analysis, the independent variables are also evaluated as being continuous.
- The sample (generally people) must be a random sample and should be taken from the predefined population.
- The dependent variables should be calculated with respect to interval and continuous scale.
- The independent variables should be calculated with respect to interval scales.
- The distributions of all the variables should be considered as normal distribution.
- The relationships between the dependent variable and the independent variable should be linear, so that a straight line can be drawn through x-y scatter diagram of the observed points.
- The independent variables can be correlated, but when the correlation is not perfect or it is said to be near-perfect, then the situation is termed as *multicollinearity*.

A regression equation

In statistical analysis, multiple regression has two functions: (a) to summarize the data (descriptive statistics) and (b) to examine the data statistically (inferential statistics).

A line in a two-dimensional or two-variable space is defined by the equation $Y=a+bX$, where Y is the variable and can be expressed in terms of a constant a and a slope b times the X variable. The constant is also termed as the *intercept*, and the slope as the *regression coefficient* or *b coefficient*.

3.4.1 Correlational Analysis

Correlational analysis helps in determining the strength of the linear relationship between the two variables *X* and *Y*. In other words, it determines as to how strongly these two variables are correlated. Karl Pearson, in 1896, developed an index or coefficient of this association in cases where the relationship is a linear one, i.e. where the trend of the relationship can be described by a straight line.

There are other indicators for the degree of relationship between two variables where the relationship is not linear, but here you will learn only about the linear relationship. Pearson's coefficient of correlation is designated by *r*. This coefficient has two characteristics:

(*i*) **The value of *r* ranges between (–1) and (+1):** If there is no relationship at all between two variables, for example, between the price of gasoline and rainfall, then its value must be zero. On the other hand, if the relationship is perfect, which means that all the points on the scatter diagram fall on the straight line, then the value of *r* is + 1 or –1, depending on the direction of the line. Other values of *r* show an intermediate degree of relationship between the two variables.

(*ii*) **The sign of the coefficient can be positive or negative:** The sign of the coefficient is positive when the slope of the line is positive and it is negative when the slope of the line is negative. For example, if the value of *Y* increases as the value of *X* increases, then the slope will be positive and so will be the sign of the coefficient of correlation, as in the case of the relationship between the height and the weight. On the other hand, if the value of *Y* decreases as the value of *X* increases, then the slope will be negative and so will be the coefficient of correlation. For example, if we see the relationship between price and demand, as the price of the product increases, the demand for the product decreases.

Hence, the coefficient of correlation *r* can be defined as the measure of strength of the linear relationship between the two variables X and Y.

CHECK YOUR PROGRESS

15. What is the multiple regression analysis?
16. What are the characteristics of Pearson's coefficient of correlation?

3.5 TIME SERIES ANALYSIS

3.5.1 Time Series Analysis Method

The time series analysis method is quite accurate where future is expected to be similar to past. The underlying assumption in time series is that the same factors will continue to influence the future patterns of economic activity in a similar manner

NOTES

as in the past. The time series analysis method is fairly sophisticated and requires expertise.

The classical approach to analysing a time series is in terms of four distinct types of variations or separate components that influence a time series. These components are:

(i) Secular Trend
(ii) Cyclic Fluctuations
(iii) Seasonal Variation
(iv) Irregular Variation

(i) Secular trend (or simply trend) (T)

The trend is the general long-term movement in the time series value of the variable (*Y*) over a fairly long period of time. The variable (*Y*) is the factor that you consider while making evaluation for the future. It could be sales, population, crime rate, and so on.

Trend is a common word, popularly used in day-to-day conversation, such as population trends, inflation trends, birth rate brands and so on. These variables are observed over a long period of time and any changes related to time are noted and calculated and a trend of these changes is established. There are many types of trends; the series may be increasing slow or increasing fast or these may be decreasing at various rates. Some remain relatively constant and some reverse their trend from growth to decline or from decline to growth over a period of time. These changes occur as a result of general tendency of the data to increase or decrease as a result of some identifiable influences.

If a trend can be determined and the rate of change can be ascertained, then tentative estimates on the same series values into the future can be made. However, such forecasts are based upon the assumption that the conditions affecting the steady growth or decline are reasonably expected to remain unchanged in the future. A change in these conditions would affect the forecasts. As an example, a time series involving increase in population over time is illustrated in Figure 3.5.

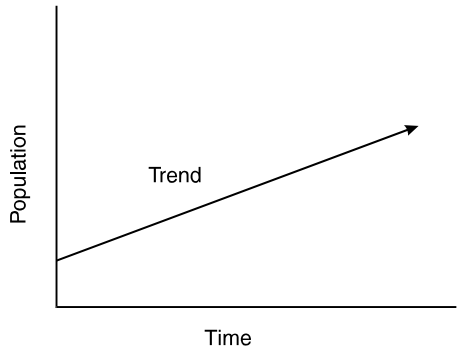


Fig. 3.5 *Time Series Showing Increase in Population over Time*

(ii) Cyclical fluctuations (C)

The cyclical fluctuations refer to regular swings or patterns that repeat over a long period of time. The movements are considered cyclical only if they occur after time intervals of more than one year. These are the changes that take place as a result of economic booms or depressions. These may be up or down and are recurrent in nature. These movements have a duration of several years, usually lasting for two to ten years. These movements also differ in intensity or amplitude and each phase of movement changes gradually into the phase that follows it. Some economists believe that the business cycle completes four phases every 12 to 15 years. These four phases are: prosperity, recession, depression and recovery. However, there is no agreement on the nature or causes of these cycles.

Even though, measurement and prediction of cyclical variation is very important for strategic planning, the reliability of such measurements is highly questionable due to the following reasons:

(*a*) Cyclic variations do not occur at regular intervals. In the twenty-five years from 1956 to 1981 in America, it is estimated that the peaks in the cyclical activity of the overall economy occurred in August 1957, April 1960, December 1969, November 1973 and January 1980.[8] This shows that they differ widely in timing, intensity and pattern, thus making reliable evaluation of trends very difficult.

(*b*) The cyclic variations are affected by many erratic, irregular and random forces which cannot be isolated and identified separately, nor can their impact be measured accurately.

The cyclic variation, for example, for revenues in an industry against time is shown in Figure 3.6.

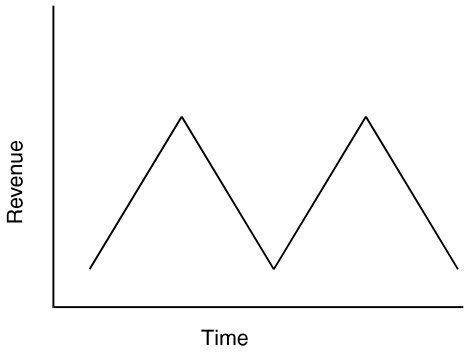


Fig. 3.6 *Cyclic Variation for Revenues Aganist Time*

[8] Berenson, Mark L. and David M. Levine. *Basic Business Statistics*. Englewood Cliff, NJ: Prentice-Hall, 1983, p. 618.

NOTES

(iii) Seasonal variations (S)

Seasonal variations involve patterns of change that repeat over a period of one year or less. Then they repeat from year to year and they are brought about by fixed events. For example, sales of consumer items increase prior to Christmas due to gift giving tradition. The sale of automobiles in America are much higher during the last 3–4 months of the year due to the introduction of new models. This data may be measured monthly or quarterly.

Since these variations repeat during a period of 12 months, they can be predicted fairly and accurately. Some factors that cause seasonal variations are:

(*a*) **Season and climate:** Changes in the climate and weather conditions have a profound effect on sales. For example, the sale of umbrellas in India is always more during monsoon or rainy season. Similarly, during winter, there is a greater demand for woollen clothes and hot drinks, while during the months of summer, there is an increase in sales of fans and air conditioners.

(*b*) **Customs and festivals:** Customs and traditions affect the pattern of seasonal spending. For example, in America, there is increase in gift sales preceding Mother's Day or Valentine's Day. In India, festivals such as Baisakhi and Diwali mean a big demand for sweets and candy. It is customary all over the world to give presents to children when they graduate from high school or college. Accordingly, the month of June, when most students graduate, is a time for the increase of sale for presents befitting the young.

(iv) Irregular (random) variations (I)

These variations are accidental, random or simply due to chance factors. Thus, they are wholly unpredictable. These fluctuations may be caused by such isolated incidents as floods, famines, strikes or wars. Sudden changes in demand or a breakthrough in a technological development may be included in this category. Accordingly, it is almost impossible to isolate and measure the value and the impact of these erratic movements on forecasting models or techniques. This phenomenon is graphically shown in Figure 3.7.

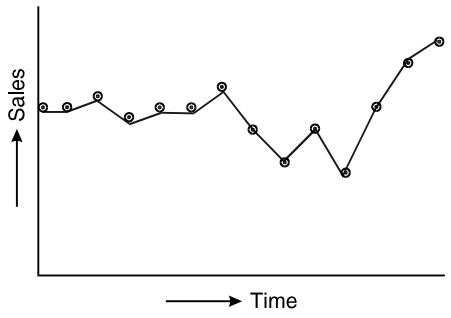


Fig. 3.7 *Irregular Variation*

It is traditionally acknowledged that the value of the time series (Y) is a function of the impact of variable trend (T), seasonal variation (S), cyclical variation (C) and irregular fluctuation (I). These relationships may vary depending upon assumptions and purposes. The effects of these four components might be additive, multiplicative, or combination thereof in a number of ways. However, the traditional time series analysis model is characterized by multiplicative relationship, so that:

$$Y = T \times S \times C \times I$$

The above model is appropriate in situations where percentage changes best represent movement in the series and the components are not viewed as absolute values but as relative values.

Another approach to define the relationship may be additive, so that:

$$Y = T + S + C + I$$

The model based on this approach is useful when the variations in the time series are in absolute values and can be separated and traced to each of these four parts and each part can be measured independently.

3.5.2 Smoothing Techniques

Smoothing techniques improve the forecasts of future trends provided that the time series are fairly stable with no significant trend, cyclical or seasonal effect and the objective is to *smooth out* the irregular component of the time series through the averaging process. There are two techniques that are generally employed for such smoothing.

(i) Moving averages

The concept of the moving averages is based on the idea that any large irregular component of time series at any point in time will have a less significant impact on the trend, if the observation at that point in time is averaged with such values immediately before and after the observation under consideration. For example, if you are interested in computing the three-period moving average for any time period, then you need to take the average of the value in such time period, the value in the period immediately preceding it and the value in the time period immediately following it. You will understand this concept better with the help of the following example.

Consider the following table which represents the number of cars sold in the first 6 weeks of the first two months of the year by a given dealer. Your objective is to calculate a 3 week moving average.

NOTES

Cars Sold in the First Two Months of the Year

Week	*Sales*
1	20
2	24
3	22
4	26
5	21
6	22

The moving average for the first 3 week period is given as follows:

$$\text{Moving average} = \frac{20 + 24 + 22}{3} = \frac{66}{3} = 22$$

This moving average can then be used to forecast the sale of cars for week 4. Since the actual number of cars sold in week 4 is 26, the error in the forecast is (26 – 22) = 4.

The calculation for the moving average for the next 3 periods is done by adding the value for week 4 and dropping the value for week 1, and taking the average for weeks 2, 3 and 4. Hence,

$$\text{Moving average} = \frac{24 + 22 + 26}{3} = \frac{72}{3} = 24$$

Then, this is considered to be the forecast of sales for week 5. Since the actual value of the sales for week 5 is 21, you have an error in your forecast of (21 – 24) = – (3).

The next moving average for weeks 3 to 5, as a forecast for week 6 is given as:

$$\text{Moving average} = \frac{22 + 26 + 21}{3} = \frac{69}{3} = 23$$

The error between the actual and the forecast value for week 6 is (22 – 23) = – (1). (Since the actual value of the sales for week 7 is not given, there is no need to forecast such values).

Your objective is to predict the trend and forecast the value of a given variable in the future as accurately as possible so that the forecast is reasonably free from random variations. To do that, you must have the sum of individual errors, as discussed above, as little as possible. However, since errors are irregular and random, it is expected that some errors would be positive in value and others negative, so that the sum of these errors would be highly distorted and would be closer to zero. This difficulty can be avoided by squaring each of the individual forecast errors and then taking the average. Naturally, the minimum values of these errors would also result in the minimum value of the 'average of the sum of squared errors'. This is shown as follows:

Week	*Time Series Value*	*Moving Average*	*Error*	*Error Squared*
1	20			
2	24			
3	22			
4	26	22	4	16
5	21	24	– 3	9
6	22	23	– 1	1

Then the average of the sum of squared errors, also known as *mean squared error* (MSE), is given as:

$$\text{MSE} = \frac{16+9+1}{3} = \frac{26}{3} = 8.67$$

The value of MSE is an often-used measure of the accuracy of the forecasting method, and the method which results in the least value of MSE is considered more accurate than others. The value of MSE can be manipulated by varying the number of data values to be included in the moving average. For example, if we had calculated the value of MSE by taking 4 periods into consideration for calculating the moving average, rather than 3, then the value of MSE would be less. Accordingly, by using trial and error method, the number of data values selected for use in forecasting would be such that the resulting MSE value would be minimum.

(ii) Exponential smoothing

In the moving average method, each observation in the moving average calculation receives the same weight. In other words, each value contributes equally towards the calculation of the moving average, irrespective of the number of time periods taken into consideration. In most actual situations, this is not a realistic assumption. Because of the dynamics of the environment over a period of time, it is more likely that the forecast for the next period would be closer to the most recent previous period than the more distant previous period, so that the more recent value should get more weight than the previous value and so on. The exponential smoothing technique uses the moving average with appropriate weights assigned to the values taken into consideration in order to arrive at a more accurate or smooth forecast. It takes into consideration the decreasing impact of the past time periods as you move further into the past time periods. This decreasing impact as you move down into the time period is exponentially distributed and hence the name *exponential smoothing*.

In this method, the smoothed value for period *t*, which is the weighted average of that period's actual value and the *smoothed* average from the previous period ($t - 1$), becomes the forecast for the next period (t + l). Then the exponential smoothing model for time period (t + l) can be expressed as follows:

$$F_{(t+1)} = \alpha Y_t + (1-\alpha)F_t$$

NOTES

where $F_{(t+1)}$ = The forecast of the time series for period $(t + 1)$

Y_t = Actual value of the time series in period t

α = Smoothing factor $(0 \le \alpha \le 1)$

F_t = Forecast of the time series for period t

The value of α is selected by the decision maker on the basis of degree of smoothing required. A small value of α means a greater degree of smoothing. A large value of α means very little smoothing. When $\alpha = 1$, then there is no smoothing at all, so that the forecast for the next time period is exactly the same as the actual value of times series in the current period. This can be worked out as:

$$F_{(t+1)} = \alpha Y_t + (1-\alpha)F_t$$

when $\alpha = 1$

$$F_{(t+1)} = Y_t + 0F_t = Y_t$$

The exponential smoothing approach is simple to use and once the value of α is selected, it requires only two pieces of information, namely Y_t and F_t to calculate $F_{(t+1)}$.

To begin with the exponential smoothing process, let F_t be equal to the actual value of the time series in period t, which is Y_1. Hence, the forecast for period 2 is written as follows:

$$F_2 = \alpha Y_1 + (1-\alpha)\, F_1$$

But since you have put $F_1 = Y_1$, hence,

$$F_2 = \alpha Y_1 + (1-\alpha)Y_1$$
$$= Y_1$$

Now you apply the exponential smoothing method to the problem of forecasting car sales as discussed in the case of moving averages. The data once again is given as follows:

Week	*Time Series Value (Y_t)*
1	20
2	24
3	22
4	26
5	21
6	22

Let $\alpha = 0.4$

Since F_2 is calculated above as equal to $Y_1 = 20$, you can calculate the value of F_3 as follows:

$$F_3 = 0.4Y_2 + (1-0.4)\, F_2$$

Since $F_2 = Y_1$

we get $F_3 = 0.4(24) + 0.6(20) = 9.6 + 12$

$= 21.6$

NOTES

Similar values can be calculated for subsequent periods, so that:

$$F_4 = 0.4Y_3 + 0.6F_3$$
$$= 0.4(22) + 0.6(21.6)$$
$$= 8.8 + 12.96$$
$$= 21.76$$
$$F_5 = 0.4Y_4 + 0.6F_4$$
$$= 0.4(26) + 0.6(21.76)$$
$$= 10.4 + 13.056$$
$$= 23.456$$
$$F_6 = 0.4Y_5 + 0.6F_5$$
$$= 0.4(21) + 0.6(23.456)$$
$$= 8.4 + 14.07$$
$$= 22.47$$

and,
$$F_7 = 0.4Y_6 + 0.6F_6$$
$$= 0.4(22) + 0.6(22.47)$$
$$= 8.8 + 13.48$$
$$= 22.28$$

Now you can compare the exponential smoothing forecast value with the actual values for the six time periods and calculate the forecast error.

Week	*Time Series Value* (Y_t)	*Exponential Smoothing Forecast Value* (F_t)	*Error* ($Y_t - F_t$)
1	20	–	–
2	24	20.000	4.0
3	22	21.600	0.4
4	26	21.760	4.24
5	21	23.456	– 2.456
6	22	22.470	– 0.47

(The value of F_7 is not considered because the value of Y_7 is not given).

Now calculate the value of MSE for this method with selected value of $\alpha = 0.4$. From the previous table:

Forecast Error	*Squared Forecast Error*
$(Y_t - F_t)$	$(Y_t - F_t)^2$
0.4	0.16
4.24	17.98
– 2.456	6.03
– 0.47	0.22
	Total = 40.39

Then,

$$\text{MSE} = 40.39/5$$
$$= 8.08$$

The previous value of MSE was 8.67. Hence, the current approach is a better one.

The choice of the value for α is very significant. Look at the exponential smoothing model again.

$$F_{(t+1)} = \alpha Y_t + (1-\alpha)F_t$$

$$= \alpha Y_t + F_t - \alpha F_t$$

$$= F_t + \alpha(Y_t - F_t)$$

Where $(Y_t - F_t)$ is the forecast error in time period *t*.

The accuracy of the forecast can be improved by carefully selecting the value of α. If the time series contains substantial random variability, then a small value of α (known as smoothing factor or smoothing constant) is preferable. On the other hand, a larger value of α would be desirable for time series with relatively little random variability $(Y_t - F_t)$.

3.5.3 Measurement of Trend and Seasonal Variations

Measuring the cyclical effect

You know that cyclic variation is a pattern that repeats over time periods longer than one year. These variations are generally unpredictable in relation to the time of occurrence, duration as well as amplitude. However, these variations have to be separated and identified. The measure you use to identify cyclical variation is the *percentage of trend* and the procedure used is known as the *residual trend*.

You have already learned that there are four components of time series: secular trend (*T*), seasonal variation (*S*), cyclical variation (*C*) and irregular (or chance) variation (*I*). Since the time period considered for seasonal variation is less than one year, it can be excluded from the study because, when you look at time series consisting of annual data spread over many years, only the secular trend, cyclical variation and irregular variation are considered.

Since the secular trend component can be described by the trend line (usually calculated by the line of regression), you can isolate cyclical and irregular components from the trend. Furthermore, since irregular variation occurs by chance and cannot be predicted or identified accurately, it can be reasonably assumed that most of the variations in time series left unexplained by the trend component can be explained by the cyclical component. In that respect, cyclical variation can be considered as the *residual*, once other causes of variation have been identified.

Calculate the measure of cyclic variation as *percentage of trend*, take the following steps:

(i) Determine the trend line (usually by regression analysis).

(ii) Compute the trend value Y_t for each time period (*t*) under consideration.

(iii) Calculate the ratio Y/Y_t for each time period.

(iv) Multiply this ratio by 100 to get the percentage of trend, so that

$$\text{Percentage of Trend} = \left(\frac{Y}{Y_t}\right)100$$

NOTES

Example 3.17: The following is the data for energy consumption (measured in quadrillions of BTU) in the United States from 1981 to 1986 as reported in the Statistical Abstracts of the United States.

Year	*Time Period (t)*	*Annual Energy Consumption (Y)*
1981	1	74.0
1982	2	70.8
1983	3	70.5
1984	4	74.1
1985	5	74.0
1986	6	73.9

Assuming a linear trend, calculate the percentage of trend for each year (cyclical variation).

Solution: First we find the secular trend by the regression line method, which is given by

$$Y_t = a + bt$$

where $$b = \frac{n\Sigma(tY) - (\Sigma t)(\Sigma Y)}{n(\Sigma t^2) - (\Sigma t)^2}$$

and $$a = \bar{Y} - b\bar{t}$$

Let us make a table for these values.

t	Y	tY	t^2
1	74.0	74.0	1
2	70.8	141.6	4
3	70.5	211.5	9
4	74.1	296.4	16
5	74.0	370.0	25
6	73.9	443.4	36
$\Sigma t = 21$	$\Sigma Y = 437.3$	$\Sigma tY = 1536.9$	$\Sigma t^2 = 91$

Substituting these values, you get

$$b = \frac{6(1536.9) - (21)(437.3)}{6(91) - (21)^2}$$

NOTES

$$= \frac{9221.4 - 9183.3}{546 - 441}$$

$$= \frac{38.1}{105} = 0.363$$

and $a = \bar{Y} - b\bar{t}$

where $\bar{Y} = \frac{\Sigma Y}{n} = \frac{437.3}{6} = 72.88$

$$\bar{t} = \frac{21}{6} = 3.5$$

Hence $a = 72.88 - 0.363(3.5)$

$= 72.88 - 1.27$

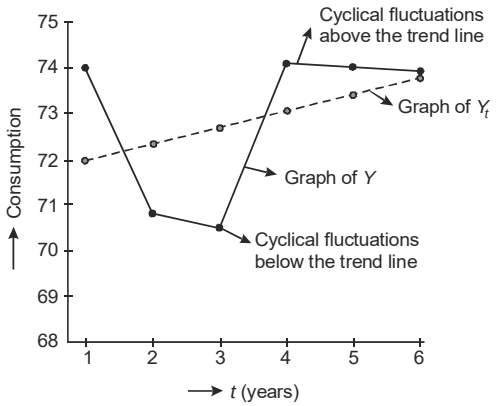
$= 71.61$

Then $Y_t = 71.61 + 0.363t$

Calculating the value of Y_t for each time period, you get the following table for percentage of trend $(Y/Y_t)100$.

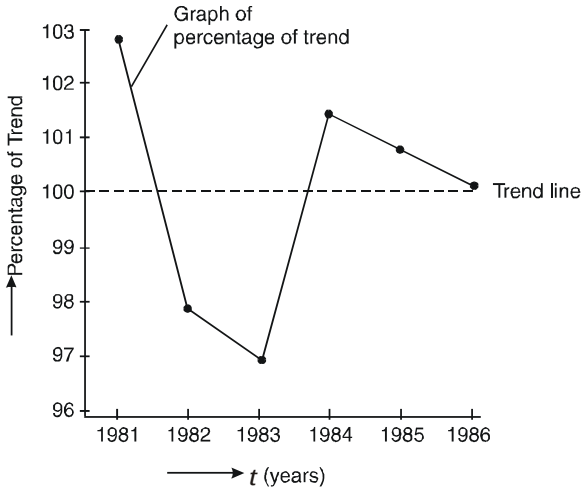
Time Period (t)	*Energy Consumption (Y)*	*Trend Line (Y_t)*	*Percentage of Trend (Y/Y_t)100*
1	74.0	71.97	102.82
2	70.8	72.34	97.87
3	70.5	72.70	96.97
4	74.1	73.06	101.42
5	74.0	73.43	100.77
6	73.9	73.79	100.15

The following graph shows the actual energy consumption (*Y*), trend line (Y_t) and the cyclical fluctuations above and below the trend line over the time period (t) for 6 years.



Frequently, you draw a graph of cyclic variation as the percentage of trend. This process eliminates the trend line and isolates the cyclical component of the time series.

It must be understood that cyclical fluctuations are not accurately predictable, and hence you cannot predict the future cyclic variations based upon such past cyclic variations.



The percentage of trend figures show that in 1981, the actual consumption of energy was 102.82 per cent of expected consumption that year and in 1983, the actual consumption was 96.97 *per cent* of the expected consumption.

Measuring seasonal variations

A seasonal variation has been defined as a predictable and repetitive movement around the trend line in a period of one year or less. For the measurement of the seasonal variation, the time interval involved may be in terms of days, weeks, months or quarters. Because of the predictability of seasonal trends, you can plan in advance to meet these variations. For example, the study of seasonal variations in the production data makes it possible to plan for hiring additional personnel for peak periods of production to accumulate an inventory of raw materials, to allocate vacation time to personnel, and so on.

In order to isolate and identify seasonal variations, you need to first eliminate, as far as possible the effects of trend, cyclical variations and irregular fluctuations on the time series. The two popular methods used for the measurement of seasonal variations are described as follows:

(i) Simple average method

This is the simplest method of isolating seasonal fluctuations in time series. It is based on the assumption that the series contain only the seasonal and irregular fluctuations. Assume that the time series involve monthly data over a time period of five years. Assume further that you want to find the seasonal index for the month of March. (The seasonal variation will be the same for March in every year. Seasonal index describes the degree of seasonal variation).

NOTES

Then the seasonal index for the month of March will be calculated as follows:

$$\text{Seasonal Index for March} = \left(\frac{\text{Monthly average for March}}{\text{Average of monthly averages}} \right) \times 100$$

The following steps can be used in the calculation of seasonal index (variation) for the month of March (or any month), over the five years period, regarding the sale of cars by one distributor.

(i) Calculate the average sale of cars for the month of March over the last five years.

(ii) Calculate the average sale of cars for each month over the five years and then calculate the average of these monthly averages.

(iii) Use the above formula to calculate seasonal index for March.

Let us say that the average sale of cars for the month of March over the period of five years is 360, and the average of all monthly average is 316. Then the seasonal index for March = (360/316)×100 = 113.92.

(ii) Ratio to moving average method

This is the most widely used method of measuring seasonal variations. The seasonal index is based upon a mean of 100 with the degree of seasonal variation (seasonal index) measured by variations away from this base value. For example, if you look at the seasonality of rental of row boats at the lake during the three summer months (a quarter) and you find that the seasonal index is 135 and you also know that the total boat rentals for the entire last year was 1680, then you can estimate the number of summer rentals for the row boats.

The average number of quarterly boats rented = 1680/4 = 420

The seasonal index, 135 for the summer quarter means that the summer rentals are 135 per cent of the average quarterly rentals.

Hence, summer rentals = 420 × (135/100) = 567

The steps required to compute the seasonal index can be enumerated by illustrating an example.

Example 3.18: Assume that a record of rental of row boats for the last three years on a quarterly basis is given as follows:

Year	*Rentals per Quarter*				*Total*
	I	II	III	IV	
1991	350	300	450	400	1500
1992	330	360	500	410	1600
1993	370	350	520	440	1680

Calculate seasonal variations using the ratio to moving average method.

Solution:

Step 1. The first step is to calculate the four quarter moving total for time series. This total is associated with the middle data point in the set of values for the four quarters, as shown below.

Year	*Quarters*	*Rentals*	*Moving Total*
1991	I	350	
	II	300	
			1500
	III	450	
	IV	400	

The moving total for the given values of four quarters is 1500 which is simply the addition of the four quarter values. This value of 1500 is placed in the middle of values 300 and 450 and recorded in the next column. For the next moving total of the four quarters, you have to drop the value of the first quarter, which is 350, from the total and add the value of the fifth quarter (in other words, first quarter of the next year), and this total will be placed in the middle of the next two values, which are 450 and 400, and so on. These values of the moving totals are shown in column 4 of the next table.

Step 2. The next step is to calculate the quarter moving average. This can be done by dividing the four quarter moving total, as calculated in Step 1, by 4, since there are four quarters. The quarters moving average is recorded in column 5 in the table. The entire table of calculations is shown below:

Year	*Quarters*	*Rentals*	*Quarter Moving Total*	*Quarter Moving Average*	*Quarter Centered Moving Average*	*Percentage of Actual to Centered Moving Average*
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1991	I	350				
	II	300				
			1500	375.0		
	III	450			372.50	120.80
			1480	370.0		
	IV	400			377.50	105.96
			1540	385.0		
1992	I	330			391.25	84.35
			1590	397.5		
	II	360			398.75	90.28
			1600	400.0		
	III	500			405.00	123.45
			1640	410.0		
	IV	410			408,75	100.30
			1630	407.5		

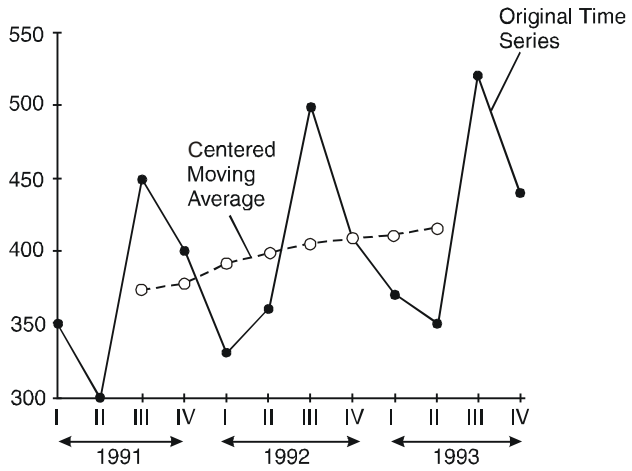
NOTES

1993	I	370			410.00	90.24
			1650	412.5		
	II	350			416.25	84.08
			1680	420.0		
	III	520				
	IV	440				

Step 3. After the moving averages for each consecutive four quarters have been taken, then you have to centre these moving averages. As you see from the above table, the quarterly moving average falls between the quarters. This is because the number of quarters is even which is 4. If you had odd number of time periods, such as 7 days of the week, then the moving average would already be centered and the third step here would not be necessary. Accordingly, you centre your averages in order to associate each average with the corresponding quarter, rather than between the quarters. This is shown in column 6, where the centered moving average is calculated as the average of the two consecutive moving averages.

The moving average (or the centered moving average) aims to eliminate seasonal and irregular fluctuations (*S* and *I*) from the original time series, so that this average represents the cyclical and trend components of the series.

As the graph for this data shows the centered moving average has smoothed the peaks and troughs of the original time series.



Step 4. Column 7 in the table contains calculated entries which are percentages of the actual values to the corresponding centered moving average values. For example, the first four quarters centered moving average of 372.50 in the table has the corresponding actual value of 450, so that the percentage of actual value to centered moving average would be:

$$\frac{\text{Actual value}}{\text{Centered moving average value}} \times 100$$

$$= \frac{450}{372.5} \times 100$$

$$= 120.80$$

NOTES

Step 5. The purpose of this step is to eliminate the remaining cyclical and irregular fluctuations still present in the values in column 7 of the table. This can be done by calculating the 'modified mean' for each quarter. The modified mean for each quarter of the three years time period under consideration, is calculated as follows.

(a) Make a table of values in column 7 of the table given on the percentage of actual to moving average values for each quarter of the three years as shown in the table.

Year	*Quarter I*	*Quarter II*	*Quarter III*	*Quarter IV*
1991	–	–	120.80	105.96
1992	84.35	90.28	123.45	100.30
1993	90.24	84.08	–	–

(b) You take the average of these values for each quarter. It should be noted that if there are many years and quarters taken into consideration instead of 3 years as you have taken, then the highest and lowest values from each quarterly data would be discarded and the average of the remaining data would be considered. By discarding the highest and lowest values from each quarter data, you tend to reduce the extreme cyclical and irregular fluctuations, which are further smoothed when you average the remaining values. Thus, the modified mean can be considered as an index of seasonal component. This modified mean for each quarter data is shown below:

$$\text{Quarter I} = \frac{84.35 + 90.24}{2} = 87.295$$

$$\text{Quarter II} = \frac{90.28 + 84.08}{2} = 87.180$$

$$\text{Quarter III} = \frac{120.80 + 123.45}{2} = 122.125$$

$$\text{Quarter IV} = \frac{105.96 + 100.30}{2} = 103.13$$

$$\text{Total} = 399.73$$

Modified means are preliminary seasonal indices. These should average 100 per cent or a total of 400 for the 4 quarters. However, our total is 399.73. This can be corrected by the following step.

NOTES

Step 6. First, you calculate an adjustment factor. This is done by dividing the desired or expected total of 400 by the actual total obtained as 399.73, so that:

$$\text{Adjustment} = \frac{400}{399.73} = 1.0007$$

By multiplying the modified mean for each quarter by the adjustment factor, you get the seasonal index for each quarter, as follows:

$$\begin{aligned}
\text{Quarter I} &= 87.295 \times 1.0007 = 87.356 \\
\text{Quarter II} &= 87.180 \times 1.0007 = 87.241 \\
\text{Quarter III} &= 122.125 \times 1.0007 = 122.201 \\
\text{Quarter IV} &= 103.13 \times 1.0007 = 103.202 \\
&\text{Total} = 400.000
\end{aligned}$$

$$\text{Average Seasonal Index} = \frac{400}{4} = 100$$

(This average seasonal index is approximated to 100 because of rounding-off errors).

The logical meaning behind this method is based on the fact that the centered moving average part of this process eliminates the influence of secular trend and cyclical fluctuations ($T \times C$). This may be represented by the following expression:

$$\frac{T \times S \times C \times I}{T \times C} = S \times I$$

Where ($T \times S \times C \times I$) is the influence of trend, seasonal variations, cyclic fluctuations and irregular or chance variations.

Thus, the ratio to moving average represents the influence of seasonal and irregular components. However, if these ratios for each quarter over a period of years are averaged, then most random or irregular fluctuations would be eliminated so that,

$$\frac{S \times I}{I} = S$$

and this would give the value of seasonal influences.

Measuring irregular variation

Typically, an irregular variation is random in nature, unpredictable and occurs over comparatively short periods of time. Because of its unpredictability, it is generally not measured or explained mathematically. Usually, subjective and logical reasoning explains such variations. For example, the current cold weather

in Brazil and Columbia is being considered responsible for increase in the price of coffee beans, because cold weather destroys coffee plants. Similarly, the Persian Gulf War, an irregular factor, resulted in increase in airline and ship travel for a number of months because of the movement of personnel and supplies. However, the irregular component can be isolated by eliminating other components from the time series data. For example, time series data contains ($T \times S \times C \times I$) components and if you can eliminate ($T \times S \times C$) elements from the data, then you are left with (I) component. You can follow the previous example to determine the (I) component as follows. The data presented below has already been calculated earlier.

Year	*Quarters*	*Rentals Time Series Values ($T \times S \times C \times I$)*	*Centered Moving Average ($T \times C$)*	*$T \times S \times C \times I$ /($T \times C$) = $S \times I$*
1991	I	350	–	–
	II	300	–	–
	III	450	372.50	1.208
	IV	400	377.50	1.060
1992	I	330	391.25	0.843
	II	360	398.75	0.903
	III	500	405.00	1.235
	IV	410	408.75	1.003
1993	I	370	410.00	0.902
	II	350	416.25	0.841
	III	520	–	–
	IV	440	–	–

The seasonal indices for each quarter have already been calculated as:

Quarter I = 87.356

Quarter II = 87.241

Quarter III = 122.201

Quarter IV = 103.202

Then the seasonal influence is given by

Quarter I = 87.356/100 = 0.874

Quarter II = 87.241/100 = 0.872

Quarter III = 122.201/100 = 1.222

Quarter IV = 103.202/100 = 1.032

The given table shows the ($S \times I$) values, (S) values and the values of (I) calculated by dividing ($S \times I$) by (S).

NOTES

Year	Quarters	(S × I)	(S)	(I)
1991	I	–	–	–
	II	–	–	–
	III	1.208	1.222	0.988
	IV	1.060	1.032	1.027
1992	I	0.843	0.874	0.965
	II	0.903	0.872	1.036
	III	1.235	1.222	1.011
	IV	1.003	1.032	0.972
1993	I	0.902	0.874	1.032
	II	0.841	0.872	0.964
	III	–	–	–
	IV	–	–	–

3.5.4 Seasonal Adjustments

Many times you read about time series values as seasonally adjusted. This is accomplished by dividing the original time series values by their corresponding seasonal indices. These deseasonalized values allow more direct and equitable comparisons of values from different time periods. For example, in comparing the demands for rental row boats (example that you have evaluated earlier), it would not be equitable to compare the demand of second quarter (spring) with the demand of third quarter (summer), when the demand is traditionally higher. However, these demand values can be compared when you remove the seasonal influence from these time series values.

The seasonally adjusted values for the demand of row boats in each quarter are based on the values previously calculated.

The following table shows the seasonally adjusted values:

Year	Quarters	Rentals (T × S × C × I)	Seasonal Variation (S)	Seasonally Adjusted Values	Rounded-Off Values
1991	I	350	–	–	–
	II	300	–	–	–
	III	450	1.222	368.25	368
	IV	400	1.032	387.60	388
1992	I	330	0.874	377.57	378
	II	360	0.872	412.80	413
	III	500	1.222	409.16	409
	IV	410	1.032	397.29	397
1993	I	370	0.874	423.34	423
	II	350	0.872	401.38	401
	III	520	–	–	–
	IV	440	–	–	–

The seasonally adjusted value for each quarter is calculated as:

$$\text{Seasonally adjusted value} = \frac{\text{Original value}}{\text{Seasonal index}}$$

NOTES

These calculations complete the process of separating and identifying the four components of the time series, namely, secular trend (*T*), seasonal variation (*S*), cyclic variation (*C*) and irregular variation (*I*).

3.5.5 Time Series and Forecasting

While the chance variations are difficult to identify, separate, control or predict, a more precise measurement of trends, cyclical effects and seasonal effects can be made in order to make the forecasts more reliable.

When a time series shows an upward or downward long-term linear trend, regression analysis can be used to estimate this trend and project the trends into forecasting the future values of the variables involved. While learning regression analysis, you have seen how the method of least squares could be used to find the best straight line relationship (line of best fit) between two variables. The equation for the straight line we used to describe the linear relationship between independent variable *X* and dependent variable *Y* was:

$$\hat{Y} = a + bX_i$$

where a = Intercept on the *Y*-axis and, b = Slope of the straight line.

In time series analysis, the independent variable is time, so you will use the symbol *t* in place of *X* and you will use the symbol Y_t in place of $\hat{Y}$ which you have used previously.

Hence, the equation for linear trend is given as

$$Y_t = a + bt$$

where Y_t = Forecast value of the time series in time period *t*

a = Intercept of the trend line on *Y*-axis

b = Slope of the trend line

t = Time period

As already discussed, you can calculate the values of *a* and *b* by the following formulae:

$$b = \frac{n\Sigma(tY) - (\Sigma t)(\Sigma Y)}{n(\Sigma t^2) - (\Sigma t)^2}, \text{ and } a = \bar{Y} - b\bar{t}$$

where Y = Actual value of the time series in time period *t*

n = Number of periods

$\bar{Y}$ = Average value of time series $= \frac{\Sigma Y}{n}$

$\bar{t}$ = Average value of $t = \frac{\Sigma t}{n}$

Knowing these values, you can calculate the value of Y_t.

Example 3.19: A car fleet owner has 5 cars which have been in the fleet for several different years. The manager wants to establish if there is a linear relationship between the age of the car and the repairs in hundreds of dollars for a given year. This way, he can predict the repair expenses for each year as the cars become older. The information for the repair costs he collected for the last year on these cars is given as follows:

Car #	*Age (t)*	*Repairs (Y)*
1	1	4
2	3	6
3	3	7
4	5	7
5	6	9

The manager wants to predict the repair expenses for next year for the two cars that are three year old now. Determine the repair costs.

Solution: The trend in repair costs suggests a linear relationship with the age of the car, so that the linear regression equation is given as

$$Y_t = a + bt$$

where

$$b = \frac{n\Sigma(tY) - (\Sigma t)(\Sigma Y)}{n(\Sigma t^2) - (\Sigma t)^2}$$

and

$$a = \bar{Y} - b\bar{t}$$

To calculate the various values, let us form a new table as follows:

Age of Car (t)	*Repair Cost (Y)*	*tY*	$\bar{t}^2$
1	4	4	1
3	6	18	9
3	7	21	9
5	7	35	25
6	9	54	36
Total 18	33	132	80

Knowing that $n = 5$, substitute these values to calculate the regression coefficients a and b.

Then

$$b = \frac{5(132) - (18)(33)}{5(80) - (18)^2}$$

$$= \frac{660 - 594}{400 - 324}$$

$$= \frac{66}{76} = 0.87$$

and

$$a = \bar{Y} - b\bar{t}$$

where

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{33}{5} = 6.6$$

NOTES

and $\bar{t} = \frac{t}{n} = \frac{18}{5} = 3.6$

Then, $a = 6.6 - 0.87(3.6)$

$= 6.6 - 3.13$

$= 3.47$

Hence, $Y_t = 3.47 + 0.87t$

The cars that are 3 years old now will be 4 years old next year, so that $t = 4$.

Hence, $Y_{(4)} = 3.47 + 0.87(4)$

$= 3.47 + 3.48$

$= 6.95$

Accordingly, the repair costs on each car that is 3 years old now are expected to be $695.00.

CHECK YOUR PROGRESS

17. Define the term irregular random variation.
18. What are the two methods adopted in smoothing techniques?
19. What are the methods used to measure seasonal variation?

3.6 SUMMARY

In this unit, you have learned that:

- Correlation analysis is the process of finding how accurately the line fits the observations, and if all the observations lie exactly on the line of best fit, the correlation is considered to be 1 or unity.
- The least squares method is the most widely used procedure for developing estimates of the model parameters.
- The correlation coefficient measures only the degree of linear association between two variables.
- Any conclusions about a cause-and-effect relationship must be based on the judgement of the analyst.
- Regression and correlation analyses are the techniques of studying how the variations in one series are related to variations in another series.
- The measurement of the degree of relationship between two or more variables is called correlation analysis and using the relationship between a known variable and an unknown variable to estimate the unknown one is termed as regression analysis. Thus, correlation measures the degree of

NOTES

relationship between the variables while regression analysis shows how the variables are related.

- Regression analysis involves identifying the relationship between a dependent variable and one or more independent variables. A model of the relationship is hypothesized, and estimates of the parameter values are used to develop an estimated regression equation.
- Various tests are employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable for the given values of independent variables.
- Regression analysis is the mathematical process of using observations to find the line of best fit through the data in order to make estimates and predictions about the behaviour of the variables. This line of best fit may be linear (straight) or curvilinear to some mathematical formula.
- Accurate forecasting is an essential element of planning of any organization or policy. This requires study of previous performances in order to forecast future activities.
- When a projection of the pattern of future economic activity is known and the level of future business activity is understood, the desirability of alternative course of action and selection of the optimum alternative can be examined and forecasted.

3.7 KEY TERMS

- **Correlation analysis**: It is the statistical tool used to describe the degree to which one variable is related to another.
- **Coefficient of determination:** It is a measure of the degree of linear association or correlation between two variables, one of which must be an independent variable and the other dependent variable.
- **Coefficient of correlation:** It is symbolically denoted by '*r*' and is an important measure to describe how well one variable is explained by another. It measures the degree of relationship between the two casually related variables.
- **Regression analysis:** It is a relationship used for making estimates and forecasts about the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s).
- **Scatter diagram:** It is also known as a Dot diagram and is used to represent two series with the known variables, i.e. independent variable plotted on the *X*-axis and the variable to be estimated, i.e. dependent variable to be plotted on the *Y*-axis on a graph paper for the given information.

- **Standard error of the estimate**: It is a measure developed by statisticians for measuring the reliability of the estimating equation.
- **Cyclic fluctuation:** It refers to regular swings or patterns that repeat over a long period of time, i.e. periods longer than one year.
- **Seasonal variation:** It involves patterns of change that repeat over a period of one year or less. The factors that cause seasonal variations are season and climate and customs and traditions.
- **Irregular variation:** These variations are unpredictable and can be accidental, random or simply due to chance factor.
- **Smoothing techniques:** It is used to smooth out the irregular components of the time series and improve the forecasts of future trends using the averaging process.
- **Seasonal adjustment:** It is calculated by dividing the original time series values by their corresponding seasonal indices. These deseasonalized values allow more direct and equitable comparisons of values from different time periods.

3.8 ANSWERS TO 'CHECK YOUR PROGRESS'

1. Correlation analysis is the statistical tool that is generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable. In fact, the word correlation refers to the relationship or interdependence between two variables. There are various phenomena that have relation to each other. For instance, when demand of a certain commodity increases, its price goes up and when its demand decreases its price comes down.
2. Whether correlation is positive or negative would depend upon the direction in which the variables are moving. If both variables are changing in the same direction, then correlation is said to be positive but when the variations in the two variables take place in opposite direction, the correlation is termed as negative.
3. The coefficient of correlation, symbolically denoted by *r*, is an important measure to describe how accurately one variable explains another. It measures the degree of relationship between the two casually related variables. The value of this coefficient can never be more than +1 or less than –1. Thus +1 and –1 are the limits of this coefficient.
4. There are several methods for calculating the coefficient of correlation. Some of the important ones are:

(a) Coefficient of correlation by the method of least squares
(b) Coefficient of correlation using simple regression coefficients
(c) Coefficient of correlation through the product moment method or Karl Pearson's coefficient of correlation

5. The coefficient of non-determination, devoted by k^2, is the ratio of unexplained variation to total variation in the *Y* variable related to the *X* variable. Coefficient of alienation is based on k^2 and it can be derived by taking the square root of k^2.
6. The various precautionary measures to be taken in using regression and correlation analyses are:
   (a) Estimating equation should not be used for extrapolation beyond the range of the observed data.
   (b) Misinterpretation of coefficients should be avoided.
   (c) Relationships that have no common bond should not be considered.
7. Regression analysis is an extremely useful tool especially in problems of business and industry for making predictions. For example, a banker could predict deposits on the basis of per capita income in the trading area of the bank. A marketing manager may plan his advertising expenditures on the basis of the expected effect on total stales revenue of a change in the level of advertising expenditure. Similarly, a hospital superintendent could project his need for beds on the basis of total population.
8. While using regression analysis for making predictions, the following assumptions are made:
   (a) There is an actual relationship between the dependent and independent variables.
   (b) The values of the dependent variable are random but the values of the independent variable are fixed quantities without error and are chosen by the experimenter.
   (c) There is clear indication of the direction of relationship. This means that the dependent variable is a function of the independent variable.
   (d) The conditions are the same when the regression model is being used. In other words, it simply means that the relationship has not changed since the regression equation was computed.
   (e) The analysis has been used to predict values within the range for which it is valid.
9. In simple linear regression model, a single variable is used to predict another variable on the assumption of linear relationship between the given variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.

NOTES

10. The two constants involved in the regression model are *a* and *b*, where *a* represents the *Y*-intercept and *b* indicates the slope of the regression line.
11. There are two methods to calculate the constants in regression models:
    (a) Scatter diagram method
    (b) Least squares method
12. It is the method of calculating constants in regression models that makes use of a scatter diagram or dot diagram. A scatter diagram is a diagram that represents two series with the known variable, i.e. the independent variable plotted on the *X*-axis and the variable to be estimated, i.e. the dependent variable to be plotted on the *Y*-axis.
13. The least squares method is a method to calculate the constants in regression models for fitting a line through the scatter diagram that minimizes the sum of the squared vertical deviations from the fitted line. In other words, the line to be fitted will pass through the points of the scatter diagram in such a way that the sum of the squares of the vertical deviations of these points from the line will be a minimum.
14. The standard error of estimate is a measure developed by statisticians for measuring the reliability of the estimating equation. Like the standard deviation, the standard error of estimate measures the variability or scatter of the observed values around the regression line.
15. The multiple regression analysis is a statistical tool that helps researchers to evaluate the effect of different factors on the consequences occurring at the same time. It analyses the relationship between several independent or predictor variables and a dependent variable.
16. The characteristics of Pearson's coefficient of correlation are as follows:
    - The value of *r* ranges between (–1) and (+1).
    - The sign of the coefficient can be positive or negative.
17. Those variations which are accidental, random or occur due to chance factors are known as irregular random variations.
18. The two methods adopted in smoothing techniques are:
    (a) Moving averages
    (b) Exponential smoothing
19. The methods used for measuring seasonal variation are:
    (a) Simple average method
    (b) Ratio to moving average method

NOTES

3.9 QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is the importance of correlation analysis?
2. How will you determine the coefficient of determination?
3. Explain the method to calculate the coefficient of correlation using the simple regression coefficient.
4. Describe Karl Pearson's method of measuring the coefficient of correlation.
5. What is the relationship between the coefficient of non-determination and coefficient of alienation?
6. List the basic precautions and limitations of regression and correlation analyses.
7. Define regression analysis.
8. How will you predict the value of a dependent variable?
9. Differentiate between scatter diagram and least squares methods.
10. Can the accuracy of an estimated equation be checked? Explain.
11. How is the standard error of estimate calculated?
12. What do you mean by trend analysis?
13. Differentiate between the secular trend and cyclic fluctuations.
14. How is irregular variation caused?
15. Define seasonal variation.
16. How will you measure the cyclical effect?
17. Describe the simple average method of isolating seasonal fluctuations in time series.
18. What are the ways to measure irregular variation?
19. How the seasonal adjustments are made?

Long-Answer Questions

1. Calculate the correlation coefficient and the two regression lines for the following information:

		Ages of Wives (in Years)				
		10–20	*20–30*	*30–40*	*40–50*	*Total*
Ages of	10–20	20	26	—	—	46
Husbands	20–30	8	14	37	—	59
(in	30–40	—	4	18	3	25
years)	40–50	—	—	4	6	10
	Total	28	44	59	9	140

NOTES

2. Two random variables have the regression with equations

$$3X + 2Y - 26 = 0$$

$$6X + Y - 31 = 0$$

Find the mean value of X as well as of Y and the correlation coefficient between X and Y. If the variance of X is 25, find σ_Y from the given data.

3. (a) Give one example of a pair of variables which would have
   - (*i*) An increasing relationship
   - (*ii*) No relationship
   - (*iii*) A decreasing relationship

   (b) Suppose that the general relationship between height in inches (X) and weight in kg (Y) is $Y' = 10 + 2.2\ (X)$. Consider that weights of persons of a given height are normally distributed with a dispersion measurable by $\sigma_e = 10$ kg.
   - (*i*) What would be the expected weight for a person whose height is 65 inches?
   - (*ii*) If a person whose height is 65 inches should weigh 161 kg, what value of e does this represent?
   - (*iii*) What reasons might account for the value of e for the person in case (*ii*)?
   - (*iv*) What would be the probability that someone whose height is 70 inches would weigh between 124 and 184 kg?

4. Calculate the correlation coefficient from the following results:

$$n = 10;\ \Sigma X = 140;\ \Sigma Y = 150$$

$$\Sigma(X - 10)^2 = 180;\ (Y - 15)^2 = 215$$

$$\Sigma(X - 10)\ (Y - 15)) = 60$$

5. Examine the following statements and state whether each one of the statements is true or false, assigning reasons to your answer.
   (a) If the value of the coefficient of correlation is 0.9, then this indicates that 90 per cent of the variation in the dependent variable has been explained by variation in the independent variable.
   (b) It would not be possible for a regression relationship to be significant if the value of r^2 was less than 0.50.
   (c) If a high significant relationship between the two variables X and Y is found, then this constitutes a definite proof that there is a causal relationship between these two variables.
   (d) A negative value of the 'b' coefficient in a regression relationship indicates a weaker relationship between the variables involved than would a positive value for the 'b' coefficient in a regression relationship.

NOTES

(e) If the value for the '*b*' coefficient in an estimating equation is less than 0.5, then the relationship will not be a significant one.

(f) $r^2 + k^2$ is always equal to 1. From this, it can also be inferred that $r + k$ is equal to 1

$$\begin{pmatrix} r = \text{coefficient of correlation; } r^2 = \text{coefficient of determination} \\ k = \text{coefficient of alienation; } k^2 = \text{coefficient of non-determination} \end{pmatrix}$$

6. Explain the meaning and significance of regression and correlation analysis.
7. What is a scatter diagram? How does it help in studying correlation between two variables? Explain.
8. Obtain the estimating equation by the method of least squares from the following information:

X (*Independent Variable*)	*Y* (*Dependent Variable*)
2	18
4	12
5	10
6	8
8	7
11	5

9. Calculate correlation coefficient from the following results:

$$n = 10;\ \Sigma X = 140;\ \Sigma Y = 150$$
$$\Sigma(X - 10)^2 = 180;\ \Sigma(Y - 15)^2 = 215$$
$$\Sigma(X - 10)\ (Y - 15) = 60$$

10.

Observation	*Test Score* *X*	*Sales* ('000 Rs) *Y*
1	73	450
2	78	490
3	92	570
4	61	380
5	87	540
6	81	500
7	77	480
8	70	430
9	65	410
10	82	490
Total	766	4740

NOTES

On the basis of information given in the table,

(a) graph the scatter diagram for the above data;

(b) find the regression equation $\hat{Y} = a + bX_i$ and draw the line corresponding to the equation on the scatter diagram;

(c) on the basis of calculated values of the coefficients of regression equation, analyse the relationship between test scores and sales;

(d) make an estimate about sales if the test score happens to be 75.

11. As a furniture retailer in a certain locality, you are interested in finding the relationship that might exist between the number of building permits issued in that locality in past years and the volume of your sales in those years. You accordingly collected the data for your sales (Y, in thousands of rupees) and the number of building permits issued (X, in hundreds) in the past 10 years. The results worked out are:

$$n = 10,\ \sum X = 200,\ \sum Y = 2200$$
$$\sum X^2 = 4600,\ \sum XY = 45{,}800,\ \sum Y^2 = 490{,}400$$

Answer the following:

(a) Calculate the coefficients of the regression equation.

(b) It is expected that there will be approximately 2000 building permits to be issued next year. On this basis, what level of sales can you expect next year?

(c) On the basis of the relationship you found in (a) what change one would expect in sales with an increase of 100 building permits?

(d) State your estimate of (b) in the (c) so that the level of confidence you place in it is 0.90.

12. Are the following two statements consistent? Give reasons for your answer.

(a) The regression coefficient of X on Y is 3.2.

(b) The regression coefficient of Y on X is 0.8.

13. Regression of savings (S) of a family on income (Y) may be expressed as $S = a + \frac{Y}{m}$, where 'a' and 'm' are constants. In random sample of 100 families, the variance of savings is one-quarter of the variance of incomes and the coefficient of correlation is found to be +0.4. Obtain the estimate of 'm'.

NOTES

14. The following data show the number of Lincoln Continental cars sold by a dealer in Queens during the 12 months of 1994.

Months	*Number Sold*
January	52
February	48
March	57
April	60
May	55
June	62
July	54
August	65
September	70
October	80
November	90
December	75

(a) Calculate the three month moving average for these data.

(b) Calculate the five month moving average for these data.

(c) Which one of these two moving averages is a better smoothing technique and why?

15. An economist has calculated the variable rate of return on money market funds for the last twelve months as follows:

Months	*Rate of Return (%)*
January	6.2
February	5.8
March	6.5
April	6.4
May	5.9
June	5.9
July	6.0
August	6.8
September	6.5
October	6.1
November	6.0
December	6.0

(a) Using a three-month moving average, forecast the rate of return for next January.

(b) Using exponential smoothing method and setting, $\alpha = 0.8$, forecast the rate of return for next January.

16. Rinkoo Camera Corporation has ten camera stores scattered in the five areas of the New York city. The president of the company wants to find out if there is any connection between the sales price and the sales volume of Nikon F-1 cameras in the various retail stores. He assigns different prices

of the same camera for the different stores and collects data for a thirty day period. The data is presented as follows. The sales volume is in number of units and the price is in dollars.

Stores	*Price*	*Volume*
1	550	420
2	600	400
3	625	300
4	575	400
5	600	340
6	500	440
7	450	500
8	480	460
9	550	400
10	650	310

(a) Plot the data.

(b) What effect would you expect on sales if the price of the camera in store number 7 is increased to $530?

(c) Calculate the points on the trend line for stores 4 and 7 and plot the trend line.

17. The following data represent the index of total industrial production (*Y*) for each quarter of the last four years.

Years	*Quarters*	*(Y)*
1st year	1	103.1
	2	107.2
	3	109.0
	4	102.1
2nd year	1	105.9
	2	109.7
	3	112.1
	4	106.0
3rd year	1	110.0
	2	112.6
	3	112.8
	4	104.3
4th year	1	107.0
	2	105.2
	3	104.8
	4	99.6

Calculate the quarter moving averages, quarter centred moving averages and percentage of actual to centred moving averages.

NOTES

18. The following data represent the values of percentage of actual data to centred moving average for each quarter of the last six years for sales of 52 inch screen Sony television of a big appliance store.

		Quarters		
Years	*Fall*	*Winter*	*Spring*	*Summer*
1990	–	–	90	30
1991	150	120	98	35
1992	160	115	95	30
1993	152	108	100	28
1994	145	115	107	32
1995	152	120	–	–

Determine the seasonal index for each quarter.

19. The Pacific Amusement Park, located in Silicon Valley, has provided the following data on the number of visitors (in thousands of admissions) during the Park's Open Seasons of Spring, Summer and Fall.

Years	*Spring*	*Summer*	*Fall*
1991	280	610	220
1992	300	725	180
1993	140	600	200
1994	200	580	180

Calculate the seasonal indices for the given data.

20. The Department of Health has compiled data on the liquor sales in the United States (in billions of dollars) for each quarter of the last four years. This quarterly data are given in the following table.

Years	*Quarters*	*Sales*
1991	1	4.5
	2	4.8
	3	5.0
	4	6.0
1992	1	4.0
	2	4.4
	3	4.9
	4	5.8
1993	1	4.2
	2	4.6
	3	5.2
	4	6.1
1994	1	4.5
	2	4.6
	3	4.9
	4	5.5

(a) Using the moving average method, find the values of the combined trend and cyclic component.

(b) Find the values of the combined seasonal and irregular component.

(c) Find the values of the seasonal indices for each quarter.

(d) Find the seasonally adjusted values for the time series.

(e) Find the value of the irregular component.

3.10 FURTHER READING

Anderson, David R., Dennis J. Sweeney and Thomas A. Williams. *Essentials of Statistics for Business and Economics*. Mumbai: Thomson Learning, 2007.

Anderson, David R., Dennis J. Sweeney and Thomas A. Williams. *Quantitative Methods for Business*. Mumbai: Thomson Learning, 2005.

Bhardwaj, R.S. *Business Statistics*. New Delhi: Excel Books, 2000.

Chandan, J.S. *Business Statistics.* New Delhi: Vikas Publishing House, 2004.

Gupta, C.B. and Vijay Gupta. *An Introduction to Statistical Methods*. New Delhi: Vikas Publishing House, 2004.

Hooda. R.P. *Statistics for Business & Economics.* New Delhi: Macmillan India Ltd., 2004.

Kothari C.R. *Quantitative Techniques*. New Delhi: Vikas Publishing House, 1984.

Levin, Richard I. and David S. Rubin. *Statistics for Business*. New Delhi: Prentice Hall of India, 1990.

Monga, G.S. *Mathematics and Statistics for Economics*. New Delhi: Vikas Publishing House.

Sancheti D.C. and V.K. Kapoor. *Business Mathematics*. New Delhi: Sultan Chand & Sons.

Zameeruddin Qazi, V.K. Sharma and S.K. Bhambri. *Business Mathematics*. New Delhi: Vikas Publishing House, 2008.

UNIT 4 PROBABILITY: THEORY AND DISTRIBUTION

Structure

4.0 INTRODUCTION

In this unit, you will learn different theories of probability and will understand why probability is considered the most important tool in statistical calculations. The subject of probability in itself is a cumbersome one hence, only the basic concepts will be discussed in this unit. The word probability or chance is very commonly used in our day-to-day conversation, and terms such as possible or probable or likely all have the same meaning. Probability can be defined as 'a measure of the likelihood that a particular event will occur'. It is a numerical measure with a value between 0 and 1, where the probability of 0 indicates that the given event cannot occur and the probability of 1 assures certainty of such an occurrence. The probability theory helps a decision maker to analyse a situation and decide accordingly. The important types of probabilities, namely *a priori* and empirical probability, objective and subjective probability, are explained with the help of solved examples.

This unit also discusses the laws of addition and multiplication. The law of addition states that when two events are mutually exclusive, the probability that either of the events will occur is the sum of their separate probabilities. The law of

NOTES

multiplication is applicable when two events occur at the same time. This unit also describes Bayes' theorem of probability theory originally stated by the Reverend Thomas Bayes. It is based on the philosophy of science and tries to clarify the relationship between theory and evidence.

This unit will also introduce you to probability distribution. The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function. The probability density function of a continuous random variable is a function which can be integrated to obtain the probability that the random variable takes a value in a given interval. The binomial distribution is used in finite sampling problems where each observation is one of two possible outcomes ('success' or 'failure'). The Poisson distribution is used for modelling rates of occurrence. The exponential distribution is used to describe units that have a constant failure rate. The term 'normal distribution' refers to a particular way in which observations will tend to pile up around a particular value rather than be spread evenly across a range of values, i.e. the central limit theorem.

4.1 UNIT OBJECTIVES

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

- Understand the basic concept of probability
- Describe the important types of probabilities
- Describe events and their role in probability
- Define binomial expansion or Jacob Bernoulli's theorem
- Describe Bayes' theorem for calculating revised probabilities
- Understand the basic concepts of probability distribution
- Describe the binomial distribution on the basis of the Bernoulli process
- Describe the Poisson distribution
- Analyse Poisson distribution as an approximation of binomial distribution
- Understand exponential distribution

4.2 PROBABILITY: BASIC CONCEPTS AND APPROACHES

Probability (usually represented by the symbol *P*) may be defined as the percentage of times in which a specific outcome would happen if an event is repeated a very large number of times. In other words, the probability of the occurrence of an event is the ratio of the number of times the event occurs (or can occur) to the number of times it and all other events occur (or can occur).

The general meaning of the word 'probability' is likelihood. Where the happening of an event is certain, the probability is said to be unity, i.e. equal to 1, and where there is absolute impossibility of happening of an event, the probability is said to be zero. But in real life such cases are rare and the probability generally lies between 0 and 1. Thus, probabilities are always greater than or equal to 0 (i.e. probabilities are never negative) and are equal to or less than 1. This being so, you can say that the weight scale of probability runs from 0 to 1 and in symbolic form it can be stated as follows:

$$P \leq 1 \text{ but } \geq 0$$

Probability can be expressed either in terms of a fraction or a decimal or a percentage but generally it is expressed in decimals.

4.2.1 The Concept of Sample Space, Sample Points and Events

A *sample space* refers to the complete set of outcomes for the situation as it did or may exist. An element in a set serving as a sample space is called a *sample point*. An *event* is a statement which refers to a particular subset of a sample space for an experiment. The meaning of these three concepts can be easily understood by means of an example. Let us consider an experiment of tossing first one coin and then another. The sample space relevant to it would then consist of all the outcomes of this experiment and can be stated as follows:

$$S = [HH, HT, TH, TT]$$

It may be noted that S = [ ] is the symbol used to represent the sample space. This sample space has four outcomes or what we call sample points, namely *HH*, *HT*, *TH* and *TT*. One or more of these sample points are called an event. One event may be that both coins fall alike and this can be represented as

$$E_1 = [HH, TT; \text{alike}]$$

The word following the semicolon explains the characteristic of your interest. If E_1 be the event of your interest and E_2 be the subset of all the remaining outcomes, then you have the following equation:

$$S = E_1 + E_2$$

4.2.2 Types of Probability

Some important types of probabilities are given as follows:

(i) *A priori* probability and empirical probability

(ii) Objective probability and subjective probability

(iii) Marginal, conditional and joint probabilities

Here, we are explaining only the first two probabilities.

(i) *A priori* probability and empirical probability

As long as any outcome or sample point concerning an experiment is not affected by external factors, each outcome or sample point is equally likely to

NOTES

occur. This assumption of each sample point being equally likely to occur is known as an *a priori* assumption (the term *a priori* refers to something which is known by reason alone) and the probability of an event worked out on this assumption is known as '*a priori probability*'. Thus, *a priori* probability is one which is worked out through deduction from assumed principles. *A priori* probability is also termed as classical probability. In our example, each sample point occurs with equal frequency. E_1 occurs twice and E_2 occurs twice; thus, the probability of E_1 is

$$P(E_1) = \frac{E_1}{E_1 + E_2} = \frac{1}{2}$$

and the probability of E_2 is

$$P(E_2) = \frac{E_2}{E_1 + E_2} = \frac{1}{2}$$

These are the examples of *a priori* probabilities. One does not need to toss the two coins a large number of times in order to predict such probabilities. *A priori* probabilities are also known as mathematical probabilities and are associated with games of chance including throws of a coin or a dice. If a coin is thrown, the probability that it will be head upwards is ½ (half) since you know that the number of possible alternatives in this case is two only. Similarly, the probability of getting a five in the single throw of a dice is $^1/_6$ (one sixth), since you know that possible alternatives in this case are six only.

Empirical probability (or statistical probability) is based on recording actual experience over a period of time and computing the proportion of items that each event occurred. Empirical probability of an event may be expressed as

$$P(E) = \frac{\text{Total number of occurences of the event } E}{\text{Total number of trials}}$$

For example, if the coin has been thrown 200 times and the head coming up was noticed in this experiment 120 times, the empirical probability of head coming up would be equal to $^{120}/_{200}$ or 0.6. 'The empirical probability of an event is taken as the relative frequency of occurrence of the event when the number of observations is very large.' You can also state that if on taking a very large number *N* out of a series of cases in which an event *E* is in question, *E* happens on *pN* occasions, the probability of the event *E* is said to be *p*. According to the laws of inertia of large numbers and statistical regularity, the more the trials, the more is the chance that empirical probability would move towards nearer and nearer and finally may become equal to *a priori* or the mathematical probability.

(ii) Objective and subjective probabilities

Objective probabilities are those which are based on definite historical information, common experience (objective evidence) or some rigorous analysis, but in the case of subjective probabilities, it is the personal experience alone which becomes the basis of the probability assignment. Let us illustrate this by an example. Suppose you have a box which contains 5 black and 15 white balls. If the balls are mixed thoroughly, then you would assign an objective probability of $5/20$ of drawing a black ball and $15/20$ of drawing a white ball. Similarly, the probabilities of various events in throwing of dice or in tossing coins are the examples of objective probabilities, since they may be, and generally are, based on reliable objective evidences. However, imagine a situation wherein a businessman is trying to decide whether or not to buy a new factory and the success of the factory largely depends on whether there is a recession or not in the next four years. If a probability is assigned to the occurrence of a recession, it would be a subjective weight based on the personal experience of the businessman. Such probability constitutes an example of subjective probability. For business decision-making purposes, the subjective probabilities are frequently required and used, especially because of the fact that reliable objective evidences are not always available.

4.2.3 Addition Rule

If two events do not happen on any one occasion, then the events are known as mutually exclusive events. In other words, events are said to be mutually exclusive when only one of the events can occur on any one trial. The probabilities of these events can be added to obtain the probability that at least one of a given collection of events will occur. This is known as *the additional rule of probability.* It can be stated as follows: if an event can happen in more than one way, all ways being mutually exclusive, the probability of its happening at all is the sum of the probabilities of its happenings in the several ways. In terms of set therory, you can state this probability relationship as follows:

$$P(A \cup B \cup C)^1 = P(A) + P(B) + P(C)$$

Provided the events A, B and C are mutually exclusive, i.e. A, B and C do not intersect at any point. But if they intersect (in that case events A, B and C are not mutually exclusive), then the probability relationship will have to be modified and can be stated as follows:

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &= -P(A \cap B) - P(A \cap C) \\ &= -P(B \cap C) + P(A \cap B \cap C)^1 \end{aligned}$$

[1] $(A \cap B \cap C)$ is pronounced as A intersection B intersection C, which means A, B and C.

NOTES

Example 4.1: Find the chance of throwing a number greater than 4 with an ordinary dice.

Solution: An ordinary dice has six faces marked as 1 to 6. Numbers greater than four can be 5 or 6. These numbers are mutually exclusive which means if you throw 5 you cannot throw 6 simultaneously and if you throw 6 you cannot throw 5 at the same time. The probability of throwing 5 with a single dice is 1/6 and the probability of throwing 6 with a single dice is also 1/6. Hence, the required probability of throwing either 5 or 6 (i.e. a number greater than 4) with a single dice is:

$$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Example 4.2: From a set of 17 cards numbered as 1, 2, ..., 17, one is drawn at random. What is the probability that the card drawn bears a number which is divisible by 3 or 7?

Solution: In this example, the numbers divisible by 3 are 3, 6, 9, 12, 15 and the numbers divisible by 7 are 7 and 14 only. The card drawn at random may bear any of these numbers (which are mutually exclusive) and the probability of each of these numbers coming is 1/17. Hence, the probability that the card drawn bears a number which is divisible by 3 or 7 is:

$$= \frac{1}{17} + \frac{1}{17} + \frac{1}{17} + \frac{1}{17} + \frac{1}{17} + \frac{1}{17} + \frac{1}{17} = \frac{7}{17}$$

4.2.4 Multiplication Rule

When two or more events occur together, their joint occurrence is called a compound event. The two events occurring together and constituting a compound event may be either independent or dependent. If two events are independent (statistically), the occurrence of the one event will not affect the probability of the occurrence of the second event.[2] On the other hand two events are said to be dependent if the occurrence of one of the events affects the probability of the occurrence of the second event. For instance, if a bag contains 10 balls and one ball is drawn from it and it is not replaced back and then a second ball is drawn, the drawing of the second ball is dependent on that of the first. But if the ball drawn is replaced, the second drawing of the ball will be taken as independent of the first. Another example of dependent events involves mutually exclusive events. If events *A* and *B* are mutually exclusive, they are dependent. Given that event *A* has occurred then probability of event *B*, the occurring must be zero since the two cannot happen simultaneously.

[2] Statistical independence or dependence is different from casual independence or dependence, simply because if two events are statistically dependent upon each other, than it does not mean that one is caused by the other.

NOTES

The probability principle in the case of independent events

When two (or more) events are independent the probability of both events (or more than two events) occurring together or in succession is equal to the product of the chances of their happening separately. This can be stated as

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

This equation shows that the probability of events *A* and *B* both occurring is equal to the probability of events *A* times the probability of event *B*, if *A* and *B* are independent events. The probability of both events *A* and *B* occurring together, i.e. $P(AB)$ or $P(A \cap B)$ is also known as *joint probability* of events *A* and *B*. *The rule concerning probability applicable in such a case is known as the multiplication theorem of probability.*

To define independence statistically we sometimes need the symbol[3]:

$$P(B/A)$$

This symbol is read as 'the probability of event *B* given that event *A* has occurred'. This also indicates the *conditional probability* of event *B* given that event *A* has taken place. With independent events,

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

and similarly

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

Thus, *with two independent events, the occurrence of one event does not affect the probability of the occurrence of the second event.*

The probability principle in the case of dependent events

The probability principle in case of dependent events remains the same as in the case of independent events. If $P(A)$ is the probability of happening of an event *A* and $P(B/A)$ is the probability of happening of an event *B* given that an event *A* has happened, then the probability that the event *A* and event *B* both happened together is as under:

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

This mean that the joint probability of *A* and *B* is equal to the conditional probability of *B* given *A* times the probability of *A*.

Similarly, you can write:

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

At this point you may now introduce the concept of marginal probability also known as unconditional probability. A marginal probability refers to the probability of happening of an event not conditional on the happening of another event. For example, $P(A)$ and $P(B)$ are examples of marginal probabilities. The term marginal is possibly used because such (marginal) probabilities are found in the margins of a joint probability table.

[3] It may be noted that $P(B/A)$ does, in no way, mean the probability of event *B* divided by *A*, but the vertical line followed by *A* simply means 'given that event *A* has occurred'.

NOTES

The relationship between conditional, marginal and joint probabilities can be stated as under:

$$P(B/A) = \frac{P(AB)}{P(A)}$$

Where, $P(B/A)$ = The conditional probability of event B given that event A has happened.

$P(AB)$ = The joint probability of event A and event B happening together.

$P(A)$ = The marginal probability of the happening of event A.

Example 4.3: What is the probability of obtaining two heads in two throws of a single coin?

Solution: The probability of obtaining a head in the first throw is 1/2.

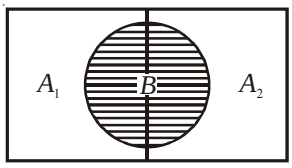
The probability of obtaining a head in the second throw is also 1/2 (it is not affected by the first throw of the coin).

The two throws (i.e. the two events) being independent, the probability of obtaining head in both of them is the product of the probability of head in the first throw and the probability of head in the second throw.

The required probability is thus: $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

4.2.5 Bayes' Theorem and their Applications

The Bayes' theorem (named after Thomas Bayes', an English Philosopher), published in 1763 in a short paper, constitutes a unique method for calculating revised probabilities. In other words, this theory concerns itself to the question of determining the probability of some even E_i given that another event A has been (or will be) observed, i.e. the theory determines the value of $P(E_i/A)$. The event A is generally thought of a sample information. As such Bayes' rule is concerned with determining the probability of an even. For example, consider a machine which is not working correctly. Given are certain sample information, say the probability of defective article is 4 per cent on the basis of sample study. The Bayes' theorem has many important applications evaluating the worth of additional information, especially in context of decision analysis. To illustrate this, let A_1 and A_2 be the sets of events which are mutually exclusive and exhaustive, so that $P(A_1) + P(A_2) = 1$, and B be a simple event which intersects each of the A events as shown in the following diagram:



NOTES

In the above diagram the part of B which is within A_1 represents the area 'A_1 and B' and the part of B within A_2 represents the area 'A_2 and B'. This being so, the probability of event A_1 given event B is:

$$P(A_1 \mid B) = \frac{P(A_1 \text{ and } B)}{P(B)}$$

And the probability of event A_2 given event B is:

$$P(A_2 \mid B) = \frac{P(A_2 \text{ and } B)}{P(B)}$$

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

$P(A_1 \text{ and } B) = P(A_1) \times P(B\,A_1)$

and $P(A_2 \text{ and } B) = P(A_2) \times P(B\,A_2)$

In general, if $A_1, A_2, A_3, \ldots, A_n$ be the set of mutually exclusive and exhaustive events, then the above expressions may be stated as follows:

$$P(A_i / B) = \frac{P(A_i \text{ and } B)}{P(B)}$$

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

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

A probability that undergoes revision through the above-stated Bayes' rule in the light of sample information is called a *posterior probability*[4]. Posterior probabilities are always conditional probabilities. You will understand this better with the help of the following example.

Example 4.4: It is not known whether a coin is fair or unfair. If the coin is fair, the probability of tail is 0.5 but if the coin is unfair the probability of a tail is 0.10. *A priori* or unconditional probability given of a fair coin is 0.80 and that of unfair coin is 0.20. The coin is tossed once and tail is the result. (*i*) What is the probability that the coin is fair? (*ii*) What is the probability that the coin is unfair?

Solution: Let the event 'fair coin' be designated by A_1 and the event 'unfair coin' by A_2. Then the given information can be put as under:

$P(A_1) = 0.80$, $P(A_2) = 0.20$ } *A priori* (or unconditional) probabilities

$P(\text{tail}/A_1) = 0.5$, $P(\text{tail}/A_2) = 0.1$ } Conditional probabilities

[4] Posterior probability is also known as inverse probability.

NOTES

$$\left.\begin{aligned} P(\text{tail and } A_1) &= P(A_1) \times P(\text{tail}/A_1) \\ &= (0.8)\,(0.5) \\ &= 0.40 \\ P(\text{tail and } A_2) &= P(A_2) \times P(\text{tail}/A_2) \\ &= (0.2)\,(0.1) \\ &= 0.02 \end{aligned}\right\} \text{ Joint probabilities}$$

A tail can occur in combination with 'fair coin' or in combination with 'unfair coin'. The probability of the former is 0.40 and of the latter is 0.02. The sum of the probabilities would result in the unconditional probability of a trail on the first toss, i.e.

$$P(\text{tail}) = 0.40 + 0.02 = 0.42$$

Thus, if a tail occurs and if it is not known whether the coin tossed once is fair coin or unfair coin, then the probability of its being a fair coin is:

$$P(A_1 \text{ tail}) = \frac{P(\text{tail and } A_1)}{P(\text{tail})} = \frac{0.40}{0.42} = 0.95$$

This is the posterior (or revised) probability of a fair coin (or A_1) given that tail is the result in the first toss of a coin obtained through the Bayes' rule.

You can similarly calculate the posterior probability of an unfair coin (or A_2) given that tail is the result in the first toss and it can be shown as follows:

$$P(A_2 \text{ tail}) = \frac{P(\text{tail and } A_2)}{P(\text{tail})} = \frac{0.02}{0.42} = 0.05$$

Thus, the revised probabilities after one toss when the toss results in tail are 0.95 of a fair coin and 0.05 of an unfair coin (initially they were 0.80 and 0.20 respectively).

4.2.6 Other Measures for Calculating Probability

In the case of simple events

Simple events are also known as single events. In such events the probability principle involved is that if an event can happen in *a* ways and fails to happen in *b* ways, all these ways being equally likely and such that not more than one of them can occur, then the probability of the event happening is $\frac{a}{(a+b)}$ and the probability of its not happening is $\frac{b}{(a+b)}$. The probability of happening of an event and the probability of its not happening must always sum to one. If the probability of happening of an event is given as $\frac{a}{(a+b)}$, then the possibility of

its not happening can be easily worked out as $1-\frac{a}{(a+b)}$ and it would be equal to $\frac{b}{(a+b)}$.

NOTES

In simplest problems on probability, the number of favourable ways and the total number of ways in which the event can happen can be counted either arithmetically or by the help of simple rules and after that the probability of happening can simply be worked out by dividing the number of favourable ways by the number of total ways in which an event can happen.

Thus, the probability of happening of an event is equal to

$$\frac{\text{Number of favourable ways}}{\text{Total number of ways in which an event can happen}}$$

Example 4.5: Find the probability that if a card is drawn at random from an ordinary pack, it is one of the court cards.

Solution: Since there are 52 cards in an ordinary pack, the total number of ways in which a card can be drawn = 52.

The number of favourable ways = 12 because King, Queen and Jack of each of the four colours is to be included in court cards.

Hence, the required probability $=\frac{12}{52}=\frac{3}{13}$

Note: Sometimes the information given may be in the form of odds in favour or odds against. If the odds are in favour of the event as *a* to *b* (or odds against *b* to *a*), then the probability of the happening of the event $=\frac{a}{(a+b)}$.

Example 4.6: Odds in favour of *A* solving the problem are 6:8. Find the probability of *A* solving the problem.

Solution: The given information means that out of 14 times, *A* can solve the problem 6 times and fails to solve the problem 8 times. Hence, the probability of solving the problem is:

$$=\frac{6}{14}=\frac{3}{7}$$

Use of binomial expansion or Jacob Bernoulli's theorem

The probability of the happening of an event in one trial being known, it is required to find the probability of its happening once, twice, thrice, ..., exactly in *n* trials. This can be done by Bernoulli's theorem or what is popularly known as the Bernoulli process. In such a case, let *p* be the probability of its happening

NOTES

and q of its failure. Then if the event will happen exactly r times in n trials, the probability must be $(r + 1)^n$ term in the expansion of $(q + p)^n$.

Since, any particular set of r trials out of n can be selected in ${}^nC_r$ ways and the chance that exactly r successes (which implies $n - r$ failures) for any of these sets is $p^r q^{n-r}$.

Therefore, the total chance is ${}^nC_r\, p^r q^{n-r}$.

Example 4.7: Three per cent of a given lot of manufactured parts are defective. What is the probability that in a sample of four items, (a) none will be defective, (b) at least one will be defective, (c) exactly two will be defective?

Solution: Let p represent the probability of defectives, q represent the probability of non-defectives and n represent the number of items in the sample. Then, the given information can be put as

$$p = 0.03$$
$$q = 0.97$$
$$n = 4$$

Using the general term of the binomial expansion, namely ${}^nC_r\, p^r q^{n-r}$

you can state as follows:

(a) The probability of non-defective part applying ${}^nC_r\, p^r q^{n-r}$

$$= {}^4C_0\, p^0 q^4$$
$$= (0.97)^4$$
$$= 0.885$$

(b) The probability of at least one defective means the probability of either 1 defective or 2 defectives or 3 defectives or 4 defectives and this can be stated as

$$= {}^4C_1 p^1 q^3 + {}^4C_2 p^2 q^2 + {}^4C_3 p^3 q^1 + {}^4C_4 p^4 q^0$$
$$= {}^4C_1 (0.03)^1\ (0.97)^3 + {}^4C_2 (0.03)^2\ (0.97)^2 + {}^4C_3 (0.03)^3\ (0.97)^1 + {}^4C_4 (0.03)^4\ (0.97)^0$$
$$= 4(0.03)(0.97)^3 + 6(0.03)^2(0.97)^2 + 4(0.03)^3(0.97) + (0.03)^4$$
$$= 0.10953076 + 0.00508086 + 0.00010476 + 0.00000081$$
$$= 0.11471719 = 0.115 \text{ approximately}$$

This can alternatively be calculated easily as follows:

The probability of nondefective = 0.885

The probability of at least one defective = [1 – Probability of nondefective]

$$= 1 - 0.885 = 0.115$$

NOTES

(c) The probability of exactly two defectives can be calculated as follows:

$$= {}^4C_2\ p^2\ q^2$$
$$= 6(0.03)^2(0.97)^2$$
$$= 0.00508086$$

Example 4.8: What is the probability of obtaining exactly three heads in seven throws with a single coin?

Solution: The probability of getting a head in a single throw:

$$p = \frac{1}{2}$$

$$\therefore \qquad q = 1 - \frac{1}{2} = \frac{1}{2}$$

You want 3 heads in seven throws with a single coin, i.e. $r = 3$ and $n = 7$. Applying ${}^nC_r p^r q^{n-r}$, you obtain the probability of obtaining exactly three heads as follows:

$$= {}^7C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^4$$

$$= 35 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{35}{128}$$

Mathematical expectation

If p happens to be the probability of the happening of an event in a single trial, then the expected number of occurrences of that event in n trials is given by np, (where n means the number of trials and p means the probability of happening of an event). Thus, the expectation may be regarded as the likely number of successes in n trials. If probability p is determined as the relative frequency in n trials, then the mathematical expectation in these n trials would be equal to the actual (observed) number of successes in these n trials. Mathematical expectation does in no way mean that the concerning event must happen the number of times given by the mathematical expectation; it simply gives the *likely number* of the happening of the event in n trials. Mathematical expectation can be illustrated by an example as follows.

Example 4.9: In a given business venture, a man can make a profit of Rs 2000 with probability 0.8 or can suffer a loss of Rs 800 with probability 0.2. Determine his expectation.

Solution: With probability 0.8, the expectation of profit

$$= \frac{8}{10} \times 2000 = \text{Rs } 1600$$

NOTES

With probability 0.2, the expectation of loss

$$= \frac{2}{10} \times 800 = \text{Rs } 160$$

His overall net expectation in the venture concerned would then clearly be Rs (1600 – 160) = Rs 1440.

Thus, in the above illustration, the concept of mathematical expectation has been extended to a discrete random variable. In the general form, you can state that if X denotes a discrete random variable which can assume the values X_1 X_2, ..., X_k with respective probabilities p_1, p_2, ..., p_k where $p_1 + p_2 + \cdots p_k = 1$, the mathematical expectation of X denoted by $E(X)$ is defined as

$$E(X) = p_1X_1 + p_2X_2 + \cdots p_kX_k$$

Markov chain

The Markov chain, also known as Markov process, refers to the sequence of experiments in which the outcome of any particular trial depends upon the outcome of the immediately preceding trial. In other words, the Markov chain is a sequence of 'states' through which a system passes at successive points in time. For example, if the system is a group of certain machines, then the 'state' of the system may be the number of machines not operating because of breakdown. Similarly, if the system is a sequence, then the 'state' of the system is the number of people waiting in the queue. If you take the various 'states' of a system as x_1, x_2, x_3, ..., x_n and consider the 'states' of the system only at discrete points in time, such as $t = 0, 1, 2, ...$, then the probability of being in a given state x_i at a given point in time t is $p_i(t)$.

In case of the simple discrete Markov process, the probability of going from state x_i at time t to state x_j at time $t + 1$ is a constant p_{ij}. This is known as the *transition probability* for the transition $x_i \rightarrow x_j$. (Remember p_{ij} is the conditional probability of the outcome x_j at a particular trial given the outcome x_i at the immediately preceding trial). If you take a sequence of states (x_i, x_j), then the probability of this particular sequence is:

$$p(x_i, x_j) = p_i(t)p_{ij}$$

If all the state probabilities are known at time 't', you can calculate them at time $t + 1$ as under:

$$p_{ij}(t+1) = \Sigma p_i(t)\, p_{ij} \quad \text{for} \quad j = 1, 2, ..., n$$

In other words, the probability of being in state x_j at time $(t + 1)$ is equal to the probability of being in state x_i at time t, i.e. times the transition probability p_{ij} summed over all the possible states of x_i.

In the form of matrix notation, the basic equation of a Markov process can be stated as under:

$$p(t+i) = Tp(t)$$

where $p(t + i)$ is the vector of state probabilities at time $(t + 1)$.

$p(t)$ is the vector of state probabilities at time t.

T is the matrix of transition probabilities whose elements come from Table 4.1 of transition probabilities.

Table 4.1 *Transition Probabilities*

From States	*To States*			
	x_1	x_2	...	x_n
x_1	p_{11}	p_{12}	...	p_{1n}
x_2	p_{21}	p_{22}	...	p_{2n}
.	.	...	...	.
.	.	...	...	.
.	.	...	...	.
.	.	...	...	.
x_n	p_{n1}	p_{n2}	...	p_{nn}

4.2.7 Probability and Venn Diagrams

You can study probability in conjunction with the set theory, where Venn diagrams can be used as a useful tool for analysing the probability related problems. However, to use Venn diagram in probability, you need to have a good understanding of it.

Understanding Venn diagram

Venn diagrams are illustrations that are used in set theory for representing all the possible logical or mathematical relationships between groups of things, which are known as sets. British philosopher and mathematician John Venn, in the year 1881, originally introduced Venn diagrams, which nowadays are considered a convenient way of illustrating the definitions included in the algebra of sets. The following example helps to understand the use of Venn diagrams in the algebra of sets.

Let us consider a universal set that contains two subsets A and B. Now you can use a Venn diagram to represent the universal sets containing the subsets A and B as shown in Figure 4.1.

NOTES

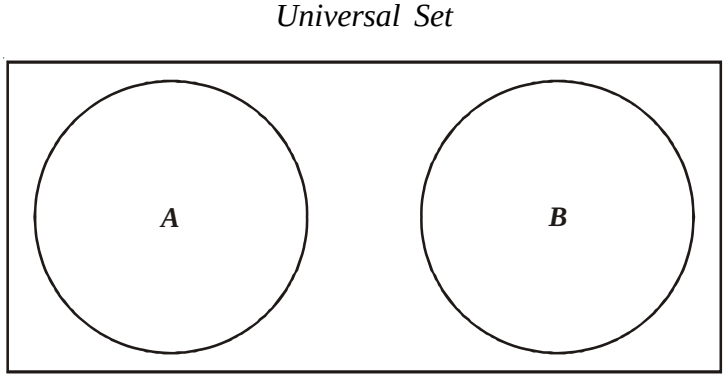


Fig. 4.1 *Venn Diagram to Represent the Universal Set with Two Subsets*

In Figure 4.1, the rectangular region represents the universal set, whereas the two circles inside the rectangular region represent the subsets of the universal set. There is no common element between the two subsets; hence, two independent circles in the Venn diagram represent them.

Now, suppose the two subsets have some elements in common. You can represent the common elements of the two subsets using a Venn diagram, where the two circles representing the subsets need to intersect and the intersecting area specifies the common elements between the two subsets. Figure 4.2 shows the Venn diagram for representing the common elements between the two subsets.

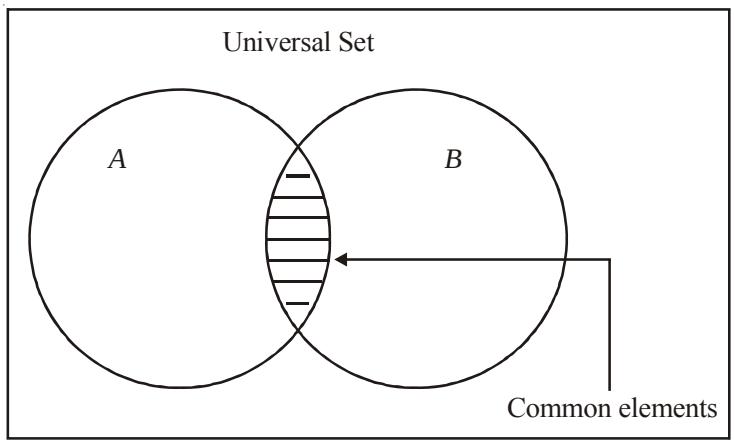


Fig. 4.2 *The Venn Diagram to Represent the Common Elements between Two Subsets*

Venn diagrams in probability

You can use Venn diagrams in probability in a similar manner as these are used in the set theory. Suppose the probability of occurrence of a certain event is $P(A)$ and the probability of occurrence of another event is $P(B)$. Therefore, $P(A \cap B)$ represents the probability of occurrence of both the events A and B, while $P(A \cup B)$ represents the probability of occurrence of events A or B. You can represent these probabilities using the Venn diagram as shown in Figure 4.3.

NOTES

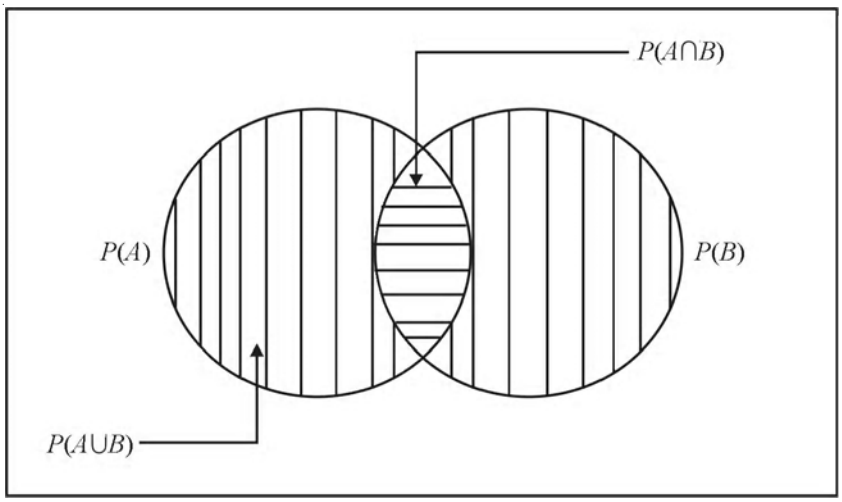


Fig. 4.3 *Venn Diagram to Represent Probability of Events*

In Figure 4.3, the rectangular area specifies the sample space, which represents the possible outcomes of all the events. The circle $P(A)$ represents the probability of occurrence of the event A; while the circle $P(B)$ represents the probability of occurrence of the event B. In the figure, the intersected portion of the two circles is represented by the horizontal lines, which specify the probability of occurrence of both A and B events. This portion of area represents the probability $P(A \cap B)$. Similarly, the parts of the two circles that are represented by vertical lines specify the probability of occurrence of A or B event. This portion of area represents the probability $P(A \cup B)$. Mathematically, the Venn diagram represents the following:

$$P(A) + P(B) - P(A \cap B) = P(A \cup B)$$

Moreover, you can also represent the probability of two mutually exclusive events using the Venn diagram. Two events are said to be mutually exclusive if they cannot occur at the same time. In other words, probability of occurrence of the two events together must be zero for two mutually exclusive events. Figure 4.4 represents two mutually exclusive events.

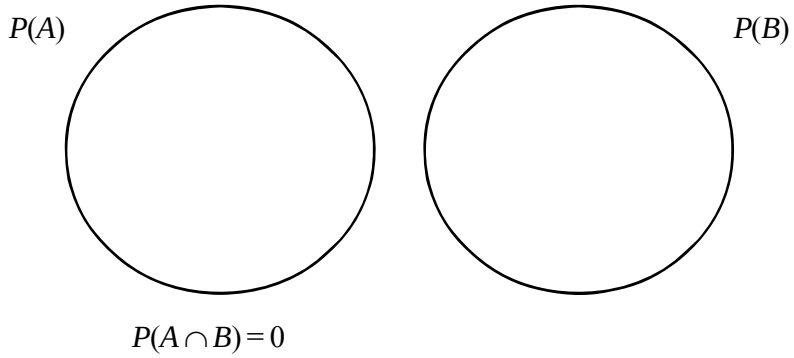


Fig. 4.4 *Venn Diagram Representing Two Mutually Exclusive Events*

NOTES

In Figure 4.4, $P(A)$ and $P(B)$ represent the probability of occurrences of two mutually exclusive events. Since both the events can not occur at the same time, the circle representing them must not overlap and therefore, from the Venn diagram it is clear that $P(A \cap B) = 0$, that is the probability of events A and B occurring together is zero.

CHECK YOUR PROGRESS

1. Define probability.
2. In what situations does one need probability theory?
3. What do you understand by sample space, sample points and events?
4. What are the various types of probabilities?
5. What is the difference between *a priori* and empirical probabilities?
6. What is the difference between objective and subjective probabilities?
7. How will you calculate the probability in the case of simple events?
8. Define a Markov chain.

4.3 PROBABILITY DISTRIBUTION

Once the random variable of interest is defined and the probabilities are assigned to all its values, it is called a probability distribution. Table 4.2 illustrates the probability distribution for various sales levels (sales level being the random variable represented as X) for a new product as stated by the sales manager:

Table 4.2 *Probability Distribution for Various Sales Levels*

Sales (in units) X_i	*Probability* *pr.* (X_i)
X_1 50	0.10
X_2 100	0.30
X_3 150	0.30
X_4 200	0.15
X_5 250	0.10
X_6 300	0.05
Total	1.00

Sometimes, the probability distribution may be presented in the form called a cumulative probability distribution. The probability distribution given in Table 4.2 can also be presented in the form of cumulative probability distribution as in Table 4.3.

NOTES

Table 4.3 *Cumulative Probability Distribution*

Sales (in units) $(X_i) \propto$	*Probability* $pr\ (X_i)$	*Cumulative Probabilities* $pr\ (X_i \leq \propto)$
X_1 50	0.10	0.10
X_2 100	0.30	0.40
X_3 150	0.30	0.70
X_4 200	0.15	0.85
X_5 250	0.10	0.95
X_6 300	0.05	1.00

The meaning of probability distribution can be made more clear if you remember the following:

- An *observed frequency distribution* (often called simply as a frequency distribution) is a listing of the observed frequencies of all the outcomes of an experiment that actually occurred while performing the experiment.
- A *probability distribution* is a listing of the probabilities of all the possible outcomes that could result if the experiment is performed. The assignment of probabilities may be based either on theoretical considerations or it may be a subjective assessment or may be based on experience.
- A *theoretical frequency distribution* is a probability distribution that describes how outcomes are expected to vary. In other words, it enlists the expected values (i.e. observed values multiplied by corresponding probabilities) of all the outcomes.

Types of probability distributions

Probability distributions can be classified as either discrete or continuous. In a *discrete probability distribution*, the variable under consideration is allowed to take only a limited number of discrete values along with corresponding probabilities. The two important discrete probability distributions are: the binomial probability distribution and the Poisson probability distribution. In a *continuous probability distribution*, the variable under consideration is allowed to take on any value within a given range. Important continuous probability distributions are exponential probability distribution and normal probability distribution.

Important disrete and continuous probability distributions are discussed later in this unit.

Probability functions

In probability distribution, it is not always necessary to calculate probabilities for each and every outcome in the sample space. There exist many mathematical formulae for many commonly encountered problems which can assign probabilities to the values of random variables. Such formulae are generally termed as probability functions. In fact, a probability function is a mathematical way of describing a given probability distribution. To select a suitable probability function

NOTES

that best fits in the given situation, you should work out the values of its parameters.[5] Once you have worked out the values of the parameters, you can then assign the probabilities, if required, using the appropriate probability function to the values of random variables. Various probability functions will be explained shortly while describing the various probability distributions.

Techniques of assigning probabilities

You can assign probability values to the random variables. Since the assignment of probabilities is not an easy task, you should observe certain rules in this context as follows:

(*i*) A probability cannot be less than zero or greater than one, i.e. $0 \leq pr \leq 1$, where *pr* represents probability.

(*ii*) The sum of all the probabilities assigned to each value of the random variable must be exactly one.

There are three techniques of assignment of probabilities to the values of the random variable:

(i) **Subjective probability assignment:** It is the technique of assigning probabilities on the basis of personal judgement. Such assignment may differ from individual to individual and depends upon the expertise of the person assigning the probabilities. It cannot be termed as a rational way of assigning probabilities but is used when the objective methods cannot be used for one reason or the other.

(ii) ***A priori* probability assignment:** It is the technique under which the probability is assigned by calculating the ratio of the number of ways in which a given outcome can occur to the total number of possible outcomes. The basic underlying assumption in using this procedure is that every possible outcome is likely to occur equally. But at times the use of this technique gives ridiculous conclusions. For example, you have to assign probability to the event that a person of age 35 will live up to age 36. There are two possible outcomes, he lives or he dies. If the probability assigned in accordance with *a priori* probability assignment is half, then the same may not represent reality. In such a situation, probability can be assigned by some other techniques.

(iii) **Empirical probability assignment:** It is an objective method of assigning probabilities and is used by the decision makers. Using this technique, the probability is assigned by calculating the relative frequency of occurrence of a given event over an infinite number of occurrences. However, in

[5] Parameters are values associated with the conditions of the experiment (or conditions of the population) from which events will occur. For example, the true average life of a product is a parameter. Usually, the constants in a function are termed as parameters.

practice, only a finite (perhaps very large) number of cases are observed and relative frequency of the event is calculated. The probability assignment through this technique may as well be unrealistic, if future conditions do not happen to be a reflection of the past.

Thus, what constitutes the 'best' method of probability assignment can only be judged in the light of what seems best to depict reality. It depends upon the nature of the problem and also on the circumstances under which the problem is being studied.

4.3.1 Binomial Distribution

Binomial distribution (or the binomial probability distribution) is a widely used probability distribution concerned with a discrete random variable and as such is an example of a discrete probability distribution. The binomial distribution describes discrete data resulting from what is often called as the Bernoulli process. The tossing of a fair coin a fixed number of times is a Bernoulli process and the outcome of such tosses can be represented by the binomial distribution. The name of Swiss mathematician Jacob Bernoulli is associated with this distribution. This distribution applies in situations where there are repeated trials of any experiment for which only one of the two mutually exclusive outcomes (often denoted as 'success' and 'failure') can result on each trial.

The Bernoulli process

Binomial distribution is considered appropriate in a Bernoulli process which has the following characteristics:

(a) **Dichotomy:** This means that each trial has only two mutually exclusive possible outcomes, e.g. 'Success' or 'failure', 'Yes' or 'No', 'Heads' or 'Tails' and the like.

(b) **Stability:** This means that the probability of the outcome of any trial is known (or given) and remains *fixed* over time, i.e. remains the same for all the trials.

(c) **Independence:** This means that the trials are statistically independent, i.e. to say the happening of an outcome or the event in any particular trial is independent of its happening in any other trial or trials.

Probability function of binomial distribution

The random variable, say X, in the binomial distribution is the number of 'successes' in n trials. The probability function of the binomial distribution is written as follows:

$$f(X = r) = {}^{n}C_{r} p^{r} q^{n-r}$$

$$r = 0, 1, 2, \ldots, n$$

NOTES

where n = Numbers of trials

p = Probability of success in a single trial

$q = (1 - p)$ = Probability of 'failure' in a single trial

r = Number of successes in 'n' trials

Parameters of binomial distribution

Binomial distribution depends upon the values of p and n which in fact are its parameters. Knowledge of p truly defines the probability of X since n is known by the definition of the problem. The probability of the happening of exactly r events in n trials can be found out using the previously stated binomial function.

The value of p also determines the general appearance of the binomial distribution, if shown graphically. In this context, the usual generalizations are as follows:

(i) When p is small (say 0.1), the binomial distribution is skewed to the right, i.e. the graph takes the form as shown in Figure 4.5.

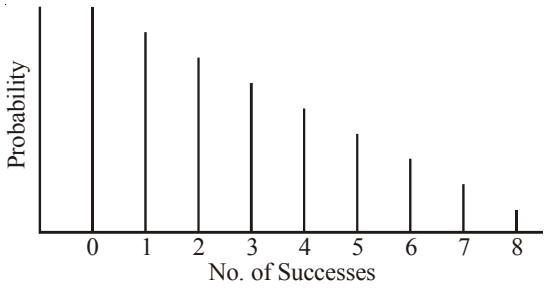


Fig. 4.5 *Bonimial Distribution Skewed to the Right*

(ii) When p is equal to 0.5, the binomial distribution is symmetrical and the graph takes the form as shown in Figure 4.6.

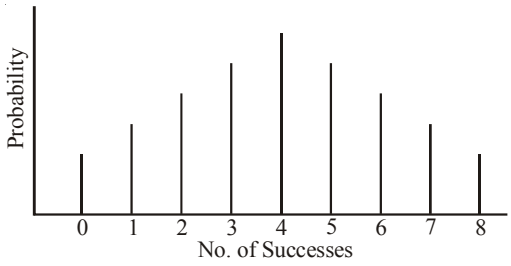


Fig. 4.6 *Symmetrical Binomial Distribution*

(iii) When p is larger than 0.5, the binomial distribution is skewed to the left and the graph takes the form as shown in Figure 4.7.

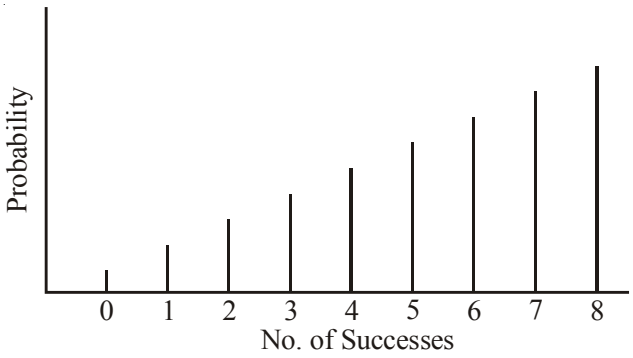


Fig. 4.7 *Bonimial Distribution Skewed to the Left*

But if '*p*' stays constant and '*n*' increases, then as '*n*' increases, the vertical lines become not only numerous but also tend to bunch up together to form a bell shape, i.e. the binomial distribution tends to become symmetrical and the graph takes the shape as shown in Figure 4.8.

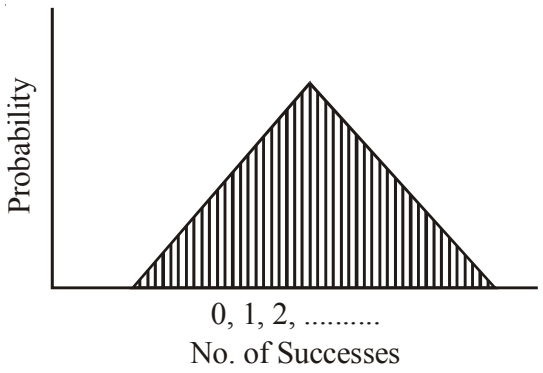


Fig. 4.8 *The Bell-Shaped Binomial Distribution*

Important measures of binomial distribution

The expected value of random variable [i.e. $E(X)$] or mean of random variable (i.e. $\bar{X}$) of the binomial distribution is equal to np and the variance of random variable is equal to npq or $np\,(1-p)$. Accordingly, the standard deviation of binomial distribution is equal to $\sqrt{npq}$. The other important measures relating to binomial distribution are as under:

$$\text{Skewness} = \frac{1-2p}{\sqrt{npq}}$$

$$\text{Kurtosis} = 3+\frac{1-6p+6q^2}{npq}$$

When to use binomial distribution

The use of binomial distribution is most appropriate in situations fulfilling the previously stated conditions. Two such situations, for example, can be described as follows:

NOTES

(i) When you have to find the probability of 6 heads in 10 throws of a fair coin.

(ii) When you have to find the probability that 3 out of 10 items produced by a machine, which produces 8 per cent defective items on an average, will be defective.

Example 4.10: A fair coin is thrown 10 times. The random variable X is the number of head(s) coming upwards. Using the binomial probability function, find the probabilities of all possible values which X can take and then verify that binomial distribution has a mean: $\bar{X} = np$ and variance: $\sigma^2 = npq$.

Solution: Since the coin is fair and so, when thrown, it can come either with head upwards or tail upwards. Hence, p (head) $= \frac{1}{2}$ and q (no head) $= \frac{1}{2}$. The required probability function is:

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

$$r = 0, 1, 2, ..., 10$$

The following table of binomial probability distribution is constructed using this function.

X_i (*Number of Heads*)	*Probability* pr_i			$X_i pr_i$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 . p_i$
0	${}^{10}C_0\, p^0\, q^{10}$	=	1/1024[6]	0/1024	–5	25	25/1024
1	${}^{10}C_1\, p^1\, q^9$	=	10/1024	10/1024	–4	16	160/1024
2	${}^{10}C_2\, p^2\, q^8$	=	45/1024	90/1024	–3	9	405/1024
3	${}^{10}C_3\, p^3\, q^7$	=	120/1024	360/1024	–2	4	480/1024
4	${}^{10}C_4\, p^4\, q^6$	=	210/1024	840/1024	–1	1	210/1024
5	${}^{10}C_5\, p^5\, q^5$	=	252/1024	1260/1024	0	0	0/1024
6	${}^{10}C_6\, p^6\, q^4$	=	210/1024	1260/1024	1	1	210/1024
7	${}^{10}C_7\, p^7\, q^3$	=	120/1024	840/1024	2	4	480/1024
8	${}^{10}C_8\, p^8\, q^2$	=	45/1024	360/1024	3	9	405/1024
9	${}^{10}C_9\, p^9\, q^1$	=	10/1024	90/1024	4	16	160/1024
10	${}^{10}C_{10}\, p^{10}\, q^0$	=	1/1024	10/1024	5	25	25/1024
				$\Sigma\bar{X} =$ 5120/1024 $\bar{X} = 5$			Variance = σ^2 = $\Sigma\ (X_i - \bar{X})^2 pr_i$ = 2560/1024 = 2.5

The mean of the binomial distribution is given by $np = 10 \times \frac{1}{2} = 5$ and the variance of this distribution is equal to $npq = 10 \times \frac{1}{2} \times \frac{1}{2} = 2.5$

These values are exactly the same as we have found them in the preceding table.

[6] The value of the binomial probability function for various values of n and p is also available in tables (known as binomial tables), which can be used to ease calculation work. The tables are of considerable help, particularly when n is large (see the Appendix).

Hence, these values stand verified with the calculated values of the two measures as shown in the table.

Fitting a binomial distribution

NOTES

When a binomial distribution is to be fitted to the given data, the following procedure is adopted:

(i) Determine the values of 'p' and 'q' keeping in view that $X = np$ and $q = (1 - p)$.

(ii) Find the probabilities for all possible values of the given random variable applying the binomial probability function, namely

$$f(X_i = r) = {}^nC_r p^r q^{n-r}$$

$$r = 0, 1, 2, \ldots, n$$

(iii) Work out the expected frequencies for all values of random variable by multiplying N (the total frequency) with the corresponding probability as worked out in case (*ii*).

The expected frequencies so calculated constitute the fitted binomial distribution to the given data.

4.3.2 Poisson Distribution

Poisson distribution is also a discrete probability distribution with which is associated the name of a Frenchman, Simeon Denis Poisson, who developed this distribution. It is frequently used in the context of operations research, and, for this reason, has a great significance for management people. It plays an important role in *queuing* theory, inventory control problems and risk models.

Unlike binomial distribution, Poisson distribution cannot be deducted on purely theoretical grounds based on the conditions of the experiment. In fact, it must be based on experience, i.e. on the empirical results of past experiments relating to the problem under study. Poisson distribution is appropriate, especially when probability of happening of an event is very small [so that q or $(1 - p)$ is almost equal to unity] and n is very large such that the average of series (namely np) is a finite number. Experience has shown that this distribution is good for calculating the probabilities associated with X occurrences in a given time period or specified area.

The random variable of interest in Poisson distribution is the number of occurrences of a given event during a given interval (interval may be time, distance, area, etc.). You use capital X to represent the discrete random variable and lower case x to represent a specific value that capital X can take. The probability function of this distribution is generally written as under:

$$f(X_i = x) = \frac{\lambda^i e^{-\lambda}}{x}$$

$$x = 0, 1, 2, \ldots$$

NOTES

where λ = Average number of occurrences per specified interval.[7] In other words, it is the mean of the distribution.

e = 2.7183 being the basis of natural logarithms.

x = Number of occurrences of a given event.

The Poisson process

The poisson distribution applies in the case of the Poisson process which has the following characteristics:

- Concerning a given random variable, the mean relating to a given interval can be estimated on the basis of past data concerning the variable under study.
- If you divide the given interval into very very small intervals you will find the following:
  (a) The probability that exactly one event will happen during the very very small interval is a very small number and is constant for every other very small interval.
  (b) The probability that two or more events will happen with in a very small interval is so small that you can assign it a zero value.
  (c) The event that happens in a given very small interval is independent, when the very small interval falls during a given interval.
  (d) The number of events in any small interval is not dependent on the number of events in any other small interval.

Parameter and important measures of poisson distribution

Poisson distribution depends upon the value of λ, the average number of occurrences per specified interval which is its only parameter. The probability of exactly x occurrences can be found out using Poisson probability function stated above.[8] The expected value or the mean of Poisson random variable is λ and its variance is also λ.[9] The standard deviation of Poisson distribution is, $\sqrt{\lambda}$.

Underlying the Poisson model is the assumption that if there are on the average λ occurrences per interval t, then there are on the average $k\lambda$ occurrences per interval kt. For example, if the number of arrivals at a service counted in

---

[7] For Binomial distribution, we had stated that mean = np

$\therefore$ Mean for Poisson distribution (or λ) = np

$\therefore$ $p = \lambda/n$

Hence, mean = $\frac{\lambda}{n}n$

[8] There are tables which give the $e^{-\lambda}$ values. These tables also give the $e^{-\lambda}\frac{\lambda^x}{x}$ values for $x = 0, 1, 2, \ldots$ for a given λ and thus facilitate the calculation work.

[9] Variance of the Binomial distributi on is npq and the variance of Poisson distribution is λ. Therefore, $\lambda = npq$. Since q is almost equal to unity and as pointed out earlier $np = \lambda$ in Poisson distribution. Hence, variance of Poisson distribution is also λ.

a given hour has a Poisson distribution with $\lambda = 4$, then y, the number of arrivals at a service counter in a given 6 hour day, has the Poisson distribution $\lambda = 24$, i.e. 6×4.

NOTES

When to use Poisson distribution

The use of Poisson distribution is resorted to in cases when you do not know the value of 'n' or when 'n' cannot be estimated with any degree of accuracy. In fact, in certain cases it does not make any sense in asking the value of 'n'. For example, if the goals scored by one team in a football match are given, it cannot be stated how many goals could not be scored. Similarly, if you watch carefully, you may find out how many times the lightning flashed but it is not possible to state how many times it did not flash. It is in such cases you use Poisson distribution. The number of deaths per day in a district in one year due to a disease, the number of scooters passing through a road per minute during a certain part of the day for a few months, the number of printing mistakes per page in a book containing many pages, etc. are a few other examples where Poisson probability distribution is generally used.

Example 4.11: Suppose that a manufactured product has 2 defects per unit of product inspected. Use Poisson distribution and calculate the probabilities of finding a product without any defect, with 3 defects and with 4 defects.

Solution: If the product has 2 defects per unit of product inspected. Hence, $\lambda = 2$.

Poisson probability function is as follows:

$$f\left(X_i = x\right) = -\frac{\lambda^x e^{-\lambda}}{x}$$

$$x = 0, 1, 2,\ldots$$

Using this probability function, you will find the required probabilities as follows:

$$P(\text{without any defects, i.e. } x = 0) = \frac{2^0 e^{-2}}{0}$$

$$= \frac{1.(0.13534)}{1} = 0.13534$$

$$P(\text{with 3 defects, i.e. } x = 3) = \frac{2^3 e^{-2}}{3} = \frac{2 \times 2 \times 2(0.13534)}{3 \times 2 \times 1}$$

$$= \frac{0.54136}{3} = 0.18045$$

$$P(\text{with 4 defects, i.e. } x = 4) = \frac{2^4 e^{-2}}{4} = \frac{2 \times 2 \times 2 \times 2(0.13534)}{4 \times 3 \times 2 \times 1}$$

$$= \frac{0.27068}{3} = 0.09023$$

NOTES

Fitting a Poisson distribution

When a Poisson distribution is to be fitted to the given data, then the following procedure is adopted:

(i) Determine the value of λ, the mean of the distribution.

(ii) Find the probabilities for all possible values of the given random variable using the Poisson probability function, namely

$$f\left(X_i = x\right) = \frac{\lambda^x . e^{-\lambda}}{x}$$

$$x = 0,\ 1,\ 2,\ldots$$

(iii) Work out the expected frequencies as follows:

$$np(X_i = x)$$

The result of case (iii) is the fitted Poisson distribution to the given data.

Poisson distribution as an approximation of binomial distribution

Under certain circumstances Poisson distribution can be considered as a reasonable approximation of binomial distribution and can be used accordingly. The circumstances which permit all this are when '*n*' is large, approaching infinity, and *p* is small, approaching zero (*n* = number of trials, *p* = probability of 'success'). Statisticians usually take the meaning of large *n*, for this purpose, when $n \geq 20$ and by small '*p*' they mean when $p \leq 0.05$. In the cases where these two conditions are fulfilled, you can use mean of the binomial distribution (namely *np*) in place of the mean of Poisson distribution (namely λ) so that the probability function of Poisson distribution becomes as follows:

$$f\left(X_i = x\right) = \frac{\left(np\right)^x e^{-(np)}}{x}$$

You can explain Poisson distribution as an approximation of the binomial distribution with the help of following example.

Example 4.12: Given are the following information:

(a) There are 20 machines in a certain factory, i.e. $n = 20$.

(b) The probability of machine going out of order during any day is 0.02.

What is the probability that exactly three machines will be out of order on the same day? Calculate the required probability using both binomial and Poissons distributions and state whether Poisson distribution is a good approximation of the binomial distribution in this case.

Solution: Probability as per Poisson probability function (using *np* in place of λ)

(since $n \geq 20$ and $p \leq 0.05$)

$$f\left(X_i = x\right) = \frac{\left(np\right)^x e^{-np}}{x}$$

NOTES

Where x means number of machines becoming out of order on the same day.

$$P(X_i = 3) = \frac{(20 \times 0.02)^3 e^{-(20 \times 0.02)}}{3}$$

$$= \frac{(0.4)^3 .(0.67032)}{3 \times 2 \times 1} = \frac{(0.064)(0.67032)}{6}$$

$$= 0.00715$$

Probability as per binomial probability function,

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

where $n = 20$, $r = 3$, $p = 0.02$ and hence $q = 0.98$

$$\therefore \quad f(X_i = 3) = {}^{20}C_3 (0.02)^3 \, (0.98)^{17}$$

$$= 0.00650$$

The difference between the probability of three machines becoming out of order on the same day calculated using probability function and binomial probability function is just 0.00065. The difference being very very small, you can state that in the given case Poisson distribution appears to be a good approximation of binomial distribution.

Example 4.13: How would you use a Poisson distribution to find approximately the probability of exactly 5 successes in 100 trials the probability of success in each trial being $p = 0.1$?

Solution: Given:

$$n = 100 \text{ and } p = 0.1$$

$$\therefore \quad \lambda = n.p = 100 \times 0.1 = 10$$

To find the required probability, the Poisson probability function can be used as an approximation to the binomial probability function as follows:

$$f(X_i = x) = \frac{\lambda^x e^{-\lambda}}{x} = \frac{(np)^x e^{-(np)}}{x}$$

or $$P(5)^7 = \frac{10^5 e^{-10}}{5} = \frac{(100000)(0.00005)}{5 \times 4 \times 3 \times 2 \times 1} = \frac{5.00000}{5 \times 4 \times 3 \times 2 \times 1}$$

$$= \frac{1}{24} = 0.042$$

4.3.3 Exponential Distribution

Exponential probability distribution is the probability distribution of time (say t), between events and as such it is continuous probability distribution concerned with the continuous random variable that takes on any value between zero and positive infinity. In the exponential distribution, you often ask the question: What is the probability that it will take x trials before the first occurrence? This

NOTES

distribution plays an important role in describing a large class of phenomena, particularly in the area of reliability theory and in *queuing* models.

The probability function of the exponential distribution is as follows:

$$F(x) = \mu e^{-}\mu^{x} \quad x \geq 0$$

where μ = The average length of the interval between two occurrences[10]

e = 2.7183 being the basis of natural logarithms

The only parameter of the exponential distribution is μ.

The expected value or mean of the exponential distribution is $1/\mu$ and its variance is $1/\mu$.[11]

The cumulative distribution (less than type) of the exponential is

$$F(x) = P(X \leq x) = 1 - e^{-}\mu^{x}, \; x \geq 0$$
$$= 0, \text{ elsewhere}$$

Thus, for instance, the probability that $x \leq 2$ is $1 - e^{-2\mu}$.

Example 4.14: In an industrial complex, the average number of fatal accidents per month is one-half. The number of accidents per month is adequately described by a Poisson distribution. What is the probability that four months will pass without a fatal accident?

Solution: You have been given that the average number of fatal accidents per month is one-half and the number of accidents per month is well described by a Poisson distribution.

Hence, $\lambda = 0.5$

$\therefore$ The average length of the time interval between two accidents $= \frac{1}{\lambda} = \frac{1}{0.5} = 2$ months, assuming exponential distribution.

Now, by using the cumulative distribution of the exponential, we can find the required probability that four months will pass without a fatal accident (i.e. $x > 4$) as follows:

$\because \quad F(x) = P(X \leq x) = 1 - e^{-}\mu^{x}$

$\therefore \quad P(X > x) = e^{-}\mu^{x}$

$\therefore \quad P(X > x) = e^{-}2^{(4)} = e^{-8} = 0.00034$

Thus, 0.00034 is the required probability that 4 months will pass without a fatal accident.

---

[10] As per the binomial probability function, this would have been worked out as follows: $p(5 \text{ successes}) = {}^{100}C_{5}(0.1)^{5}(0.9)^{95}$

[11] μ is equal to $\frac{1}{\lambda}$, where λ is the mean of Poisson distribution.

NOTES

4.3.4 Normal Distribution

Among all the probability distributions, the normal probability distribution is by far the most important and frequently used continuous probability distribution. This is so because this distribution well fits in many types of problems. This distribution is of special significance in inferential statistics since it describes probabilistically the link between a statistic and a parameter (i.e. between the sample results and the population from which the sample is drawn). The name of Karl Gauss, the eighteenth century mathematician-astronomer, is associated with this distribution and in honour of his contribution, this distribution is often known as the **Gaussian distribution**.

The normal distribution can be theoretically derived as the limiting form of many discrete distributions. For instance, if in the binomial expansion of $(p + q)^n$, the value of 'n' is infinity and $p = q = \frac{1}{2}$, then a perfectly smooth symmetrical curve would be obtained. Even if the values of p and q are not equal but if the value of the exponent 'n' happens to be very very large, you get a curve normal probability smooth and symmetrical. Such curves are called normal probability curves (or at times known as normal curves of error) and represent the normal distributions.[12]

The probability function in the case of normal probability distribution[13] is given as follows:

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

where μ = The mean of the distribution

σ^2 = Variance of the distribution

---

[12] Quite often, mathematicians use the normal approximation of the binomial distribution whenever 'n' is equal to or greater than 30 and np and nq each are greater than 5.

[13] The equation of the normal curve in its simplest form is

$$y = y_0 e^{\left(x^2/2\sigma^2\right)}$$

where y = The computed height of an ordinate at a distance of X from the mean.

y_0 = The height of the maximum ordinate at the mean. It is a constant in the equation and is worked out as follows:

$$y_0 = \frac{N_i}{\sqrt{2\pi}}$$

where N = Total number of items in the sample

i = Class interval

π = 3.1416

$\therefore$ = $\sqrt{2\pi} = \sqrt{6.2832} = 2.5066$

and e = 2.71828, base of natural logarithms

σ = Standard deviation

X = Any given value of the dependent variable expressed as a deviation from the mean

NOTES

The normal distribution is thus defined by two parameters, namely μ and σ^2. This distribution can be represented graphically (Figure 4.9).

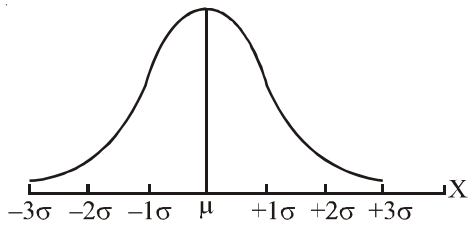


Fig. 4.9 *Curve Representing Normal Distribution*

Characteristics of normal distribution

The characteristics of the normal distribution or those of a normal curve are as follows:

(i) It is symmetric distribution.[14]

(ii) The mean μ defines where the peak of the curve occurs. In other words, the ordinate at the mean is the highest ordinate. The height of the ordinate at a distance of one standard deviation from mean is 60.653 per cent of the height of the mean ordinate, and similarly, the height of other ordinates at various standard deviations (σ_s) from mean happens to be a fixed relationship with the height of the mean ordinate.

(iii) The curve is asymptotic to the baseline which means that it continues to approach but never touches the horizontal axis.

(iv) The variance (σ^2) defines the spread of the curve.

(v) Area enclosed between mean ordinate and an ordinate at a distance of one standard deviation from the mean is always 34.134 per cent of the total area of the curve. It means that the area enclosed between two ordinates at one sigma distance (SD) from the mean on either side would always be 68.268 per cent of the total area. This is shown in Figure 4.10.

[14] A symmetric distribution is one which has no skewness. As such, it has the following statistical properties:
(a) Mean=Mode=Median (i.e. X=Z=M)
(b) (Upper Quantile – Median)=(Median – Lower Quantile) (i.e. $Q_3 - M = M - Q_1$)
(c) Mean Deviation=0.7979(Standard Deviation)
(d) $\frac{Q_3 - Q_1}{2} = 0.6745$ (Standard Deviation)

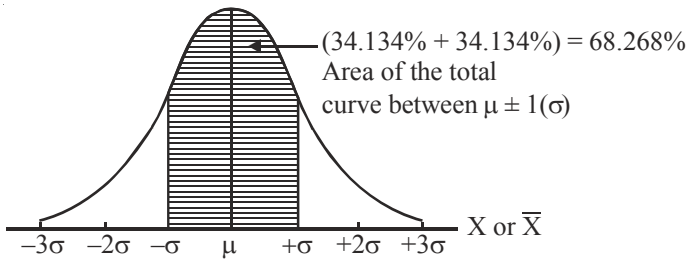


Fig. 4.10 *An Area Enclosed between Two Ordinates at One SD*

Similarly, the other area relationships are as given in Table 4.4.

Table 4.4 *Area Relationships*

Between		*Area Covered to Total Area of the Normal Curve*[15]
μ ± 1	SD	68.27%
μ ± 2	SD	95.45%
μ ± 3	SD	99.73%
μ ± 1.96	SD	95%
μ ± 2.578	SD	99%
μ ± 0.6745	SD	50%

(vi) The normal distribution has only one mode since the curve has a single peak. In other words, it is always a unimodal distribution.

(vii) The maximum ordinate divides the graph of normal curve into two equal parts.

(viii) In addition to all the above stated characteristics the curve has the following properties:

(a) $\mu = \bar{x}$

(b) $\mu_2 = \sigma^2 =$ Variance

(c) $\mu_4 = 3\sigma^4$

(d) Moment coefficient of Kurtosis = 3

Family of normal distributions or curves

You can have several normal probability distributions but each particular normal distribution is being defined by its two parameters, namely the mean (μ) and the standard deviation (σ). There is, thus, not a single normal curve but rather a family of normal curves. Figures 4.11–4.13 exhibit some of these normal curves:

[15] This also means that in a normal distribution, the probability of area lying between various limits is as follows:

Limits	*Probability of area lying within the stated limits*	
μ ± 1 SD	0.6827	
μ ± 2 SD	0.9545	
μ ± 3 SD	0.9973	(This means that almost all cases lie within μ ± 3 SD limits.)

NOTES

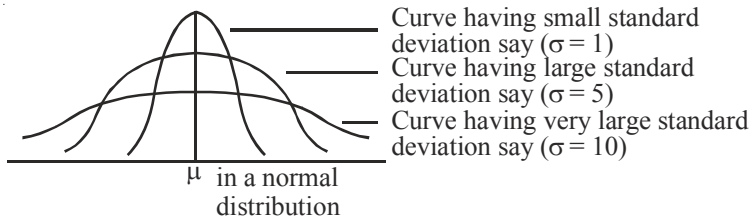


Fig. 4.11 *Normal Curves with Identical Means but Different Standard Deviations*

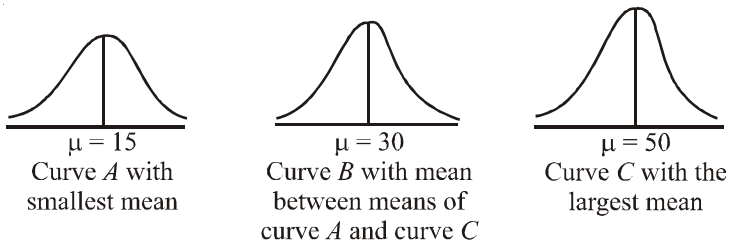


Fig. 4.12 *Normal Curves with Identical Standard Deviation but Each with Different Means*

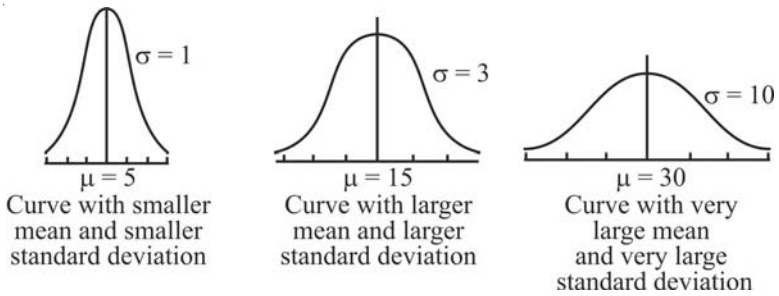


Fig. 4.13 *Normal Curves Each with Different Standard Deviations and Different Means*

How to measure the area under the normal curve?

You have learned about the the area relationships involving certain intervals of standard deviations (plus and minus) from the means that are true in case of a normal curve. But what should be done in all other cases? You can make use of the statistical tables constructed by mathematicians for the purpose. Using these tables, you can find the area (or probability, taking the entire area of the curve as equal to 1) that the normally distributed random variable will lie within certain distances from the mean. These distances are defined in terms of standard deviations. While using the tables showing the area under the normal curve, you talk in terms of standard variate (symbolically Z) which really means standard deviations without units of measurement and this 'Z' is worked out as under:

$$Z = \frac{X - \mu}{\sigma}$$

NOTES

where Z = The standard variate (or number of standard deviations from X to the mean of the distribution)

X = Value of the random variable under consideration

μ = Mean of the distribution of the random variable

σ = Standard deviation of the distribution

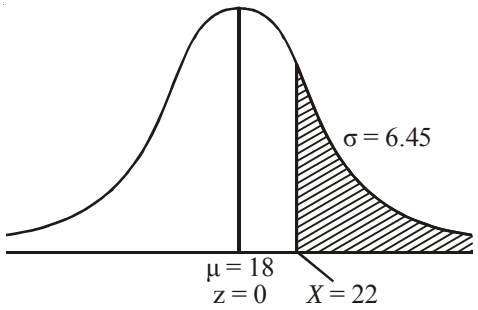
The table showing the area under the normal curve (often termed as the standard normal probability distribution table) is organized in terms of standard variate (or Z) values. It gives the values for only half the area under the normal curve, beginning with $Z = 0$ at the mean. Since the normal distribution is perfectly symmetrical, the values true for one half of the curve are also true for the other half.

Example 4.15: A banker claims that the life of a regular saving account opened with his bank averages 18 months with a standard deviation of 6.45 months. Answer the following questions:

(a) What is the probability that there will still be money in 22 months in a savings account opened with the said bank by a depositor?

(b) What is the probability that the account will have been closed before two years?

Solution: (a) For finding the required probability, you are interested in the area of the portion of the normal curve as shaded and shown in the following diagram:



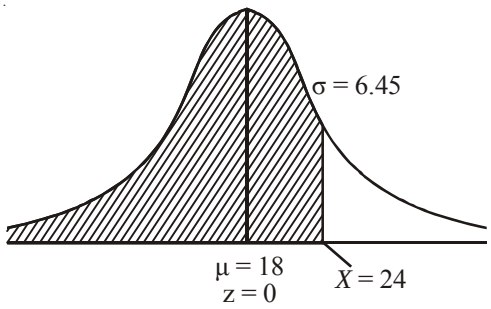
Calculate Z as under:

$$Z = \frac{X - \mu}{\sigma} = \frac{22 - 18}{6.45} = 0.62$$

The value from the table showing the area under the normal curve for $Z = 0.62$ is 0.2324. This means that the area of the curve between $\mu = 18$ and $X = 22$ is 0.2324. Hence, the area of the shaded portion of the curve is $(0.5) - (0.2324) = 0.2676$ since the area of the entire right hand portion of the curve always happens to be 0.5. Thus, the probability that there will still be money in 22 months in a savings account is 0.2676.

NOTES

(b) For finding the required probability, you are interested in the area of the portion of the normal curve as shaded and shown in the following figure:



Calculate Z as under:

$$Z = \frac{24 - 18}{6.45} = 0.93$$

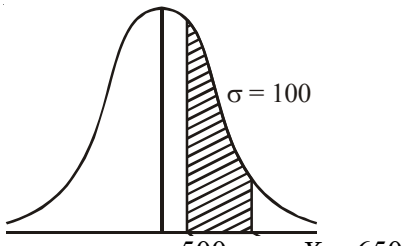
The value from the concerning table, when $Z = 0.93$, is 0.3238 which refers to the area of the curve between $\mu = 18$ and $X = 24$. The area of the entire left hand portion of the curve is 0.5 as usual.

Hence, the area of the shaded portion is (0.5) + (0.3238) = 0.8238 which is the required probability that the account will have been closed before two years, i.e. before 24 months.

Example 4.16: Regarding a certain normal distribution concerning the income of the individuals, you are given that mean=500 rupees and standard deviation =100 rupees. Find the probability that an individual selected at random will belong to income group:

(a) Rs 550 to Rs 650 (b) Rs 420 to 570

Solution: (a) For finding the required probability, you are interested in the area of the portion of the normal curve as shaded and shown in the following figure:



For finding the area of the curve between $X = 550$ to 650, do the following calculations:

$$Z = \frac{550 - 500}{100} = \frac{50}{100} = 0.50$$

NOTES

Corresponding to which the area between μ = 500 and X = 550 in the curve as per table is equal to 0.1915 and

$$Z = \frac{650 - 500}{100} = \frac{150}{100} = 1.5$$

Corresponding to which the area between μ = 500 and X = 650 in the curve as per table is equal to 0.4332

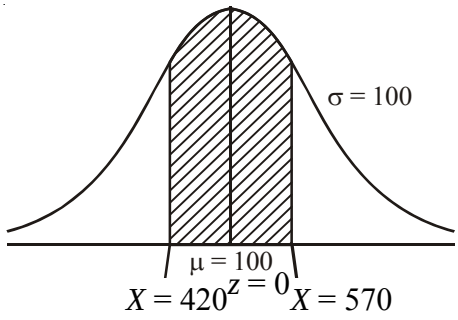
Hence, the area of the curve that lies between X = 550 and X = 650 is

$$(0.4332) - (0.1915) = 0.2417$$

This is the required probability that an individual selected at random will belong to income group of Rs 550 to Rs 650.

(b) For finding the required probability, you are interested in the area of the portion of the normal curve as shaded and shown in the following figure:

To find the area of the shaded portion we make the following calculations:



$$Z = \frac{570 - 500}{100} = 0.70$$

Corresponding to which the area between μ = 500 and X = 570 in the curve as per table is equal to 0.2580.

and $$Z = \frac{420 - 500}{100} = -0.80$$

Corresponding to which the area between μ = 500 and X = 420 in the curve as per table is equal to 0.2881.

Hence, the required area in the curve between X = 420 and X = 570 is:

$$(0.2580) + (0.2881) = 0.5461$$

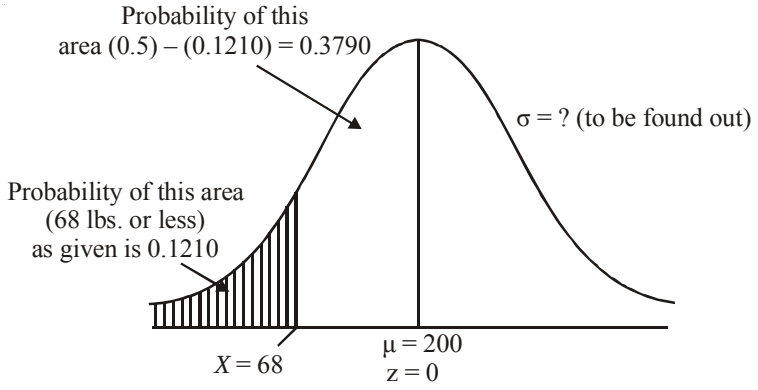
This is the required probability that an individual selected at random will belong to income group of Rs 420 to Rs 570.

Example 4.17: A certain company manufactures $1\frac{1''}{2}$ all-purpose rope made from imported hemp. The manager of the company knows that the average load-bearing capacity of the rope is 200 lbs. Assuming that normal distribution applies, find the

standard deviation of load-bearing capacity for the $1\frac{1''}{2}$ rope if it is given that the rope has a 0.1210 probability of breaking with 68 lbs or less pull.

NOTES

Solution: Given information can be depicted in a normal curve as shown in the following figure:



If the probability of the area falling within $\mu = 200$ and $X = 68$ is 0.3790 as stated above, the corresponding value of Z as per the table[16] showing the area of the normal curve is – 1.17 (minus sign indicates that we are in the left portion of the curve)

Now to find σ, you can write:

$$Z = \frac{X - \mu}{\sigma}$$

or $$-1.17 = \frac{68 - 200}{\sigma}$$

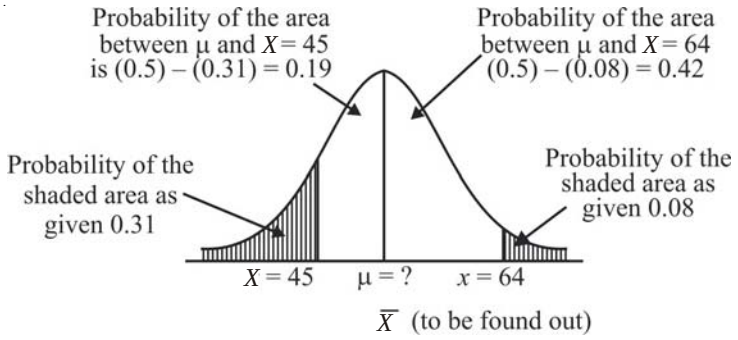
or $$-1.17\sigma = -132$$

or $$\sigma = 112.8 \text{ lbs approx.}$$

Thus, the required standard deviation is 112.8 lbs approximately.

Example 4.18: In a normal distribution, 31 per cent items are below 45 and 8 per cent are above 64. Find the $\overline{X}$ and σ of this distribution.

Solution: You can depict the given information in a normal curve as follows:



[16] The table is to be read in the reverse order for finding Z value (see the Appendix).

If the probability of the area falling within μ and $X = 45$ is 0.19 as stated above, the corresponding value of Z from the table showing the area of the normal curve is –0.50. Since, you are in the left portion of the curve so we can express this as under,

$$-0.50 = \frac{45 - \mu}{\sigma} \quad \text{(i)}$$

Similarly, if the probability of the area falling within μ and $X = 64$ is 0.42 as stated above, the corresponding value of Z from the area table is +1.41. Since, you are in the right portion of the curve, so you can express this as under:

$$1.41 = \frac{64 - \mu}{\sigma} \quad \text{(ii)}$$

If you solve equations (i) and (ii) above to obtain the value of μ or $\bar{X}$, you have:

$$-0.5\,\sigma = 45 - \mu \quad \text{(iii)}$$

$$1.41\,\sigma = 64 - \mu \quad \text{(iv)}$$

By subtracting the equation (iv) from (iii), you have:

$$-1.91\,\sigma = -19$$

$$\therefore \quad \sigma = 10$$

Putting $\sigma = 10$ in equation (iii), you have:

$$-5 = 45 - \mu$$

$$\therefore \quad \mu = 50$$

Hence, $\bar{X}$ (or μ)=50 and σ=10 for the concerning normal distribution.

Check Your Progress

9. What are the types of probability distributions?
10. Write the probability function of binomial distribution.
11. What are the different parameters of binomial distribution?
12. Under what circumstances would you use binomial distribution?
13. What is Poisson distribution?
14. What is the use of exponential distribution?
15. Define a normal distribution.

4.4 SUMMARY

In this unit, you have learned that:

- In decision-making processes, mathematical theory of probability is considered as an important and fundamental tool and is used by decision makers to evaluate the outcome of an experiment.

NOTES

- Probability is a measure of how likely an event is to happen or in other terms it can be used to estimate frequencies of outcomes in random experiments, which is always greater than or equal to 0.
- The probability of the happening of an event in one trial being known, its probability of happening in *n* trials can be computed by using Bernoulli's theorem.
- The revised probabilities can be determined by the unique method of Bayes' theorem.
- Mathematical theory of probability furnishes an important tool which can be of great help to the decision maker. As a result, probability theory has become an indispensable tool for all types of formal studies that involve uncertainty.
- Binomial distribution is probably the best known of discrete distributions. The normal distribution, or *Z*-distribution, is often used to approximate the binomial distribution.
- If the sample size is very large, the Poisson distribution is a philosophically more correct alternative to binomial distribution than normal distribution.
- One of the main differences between the Poisson distribution and the binomial distribution is that in using the binomial distribution all eligible phenomena are studied, whereas in the Poisson, only the cases with a particular outcome are studied.
- Exponential distribution is a very commonly used distribution in reliability engineering. The reason for its widespread use lies in its simplicity, so much that it has even been employed in cases to which it does not apply directly.
- Amongst all types of distributions, the normal probability distribution is by far the most important and frequently used distribution because it fits well in many types of problems.

4.5 KEY TERMS

- **Probability:** It can be defined as a measure of the likelihood that a particular event will occur.
- **Event:** It is an outcome or a set of outcomes of an activity or a result of a trial.
- **Simple event:** It is a single possible outcome of an experiment.
- **Compound event:** It is also known as a joint event, with two or more simple events in it.
- **Sample space:** It is the collection of all possible events or outcomes of an experiment.

- **Addition rule:** It states that when two events are mutually exclusive, the probability that either of the events will occur is the sum of their separate probabilities.
- **Multiplication rule:** It is applied when it is necessary to compute the probability when both events *A* and *B* occur at the same time. Different rules are applied for different conditions.
- **Venn diagram:** It refers to a diagram which represents all the possible logical or mathematical relationships between groups of things known as sets.
- **Binomial distribution:** It is also called the Bernoulli process and is used to describe a discrete random variable.
- **Poisson distribution:** It is used to describe the empirical results of past experiments relating to the problem and plays an important role in queuing theory, inventory control problems and risk models.
- **Exponential distribution:** It is a continuous probability distribution and is used to describe the probability distribution of time between two events.
- **Normal distribution:** It is referred to as most important and frequently used continuous probability distribution as it fits well in many types of problems.

4.6 ANSWERS TO 'CHECK YOUR PROGRESS'

1. Probability can be defined as the percentage of times in which a specific outcome would happen if an event were repeated a very large number of times. In other words, probability of the occurrence of an event is the ratio of the number of times the event can occur to the number of times it and all other events can occur. Probability is usually represented by the symbol *P*.
2. Probability theory is needed in situations when in the business or industrial field one needs to make predictions.
3. A sample space is the complete set of outcomes for the situation as it did or may exist, while an element in a set serving as a sample space is called a sample point. An event is a statement that refers to a particular subset of a sample space for an experiment.
4. The various types of probabilities are:
   (a) *A priori* probability and empirical probability
   (b) Objective probability and subjective probability
   (c) Marginal, conditional and joint probabilities
5. As long as any outcome or sample point concerning an experiment is not affected by external factors, each outcome or sample point is equally likely to occur. This assumption of each sample point being equally likely to occur is known as an *a priori* assumption and the probability of an event worked

NOTES

out on this assumption is known as *a priori* probability. On the other hand, empirical probability is based on recording actual experience over a period of time and computing the proportion of items that each event occurred.

6. Objective probabilities are those probabilities, which are based on definite historical information, common experience or rigorous analysis; while in the case of subjective probabilities, it is the personal experience alone that becomes the basis of the probability assignment.
7. The probability of simple events can be calculated using the following formula:

$$\text{Probability of happening of an event} = \frac{\text{Number of favourable ways}}{\text{Total number of ways in which an event can happen}}$$

8. The Markov chain, also known as the Markov process, refers to a sequence of experiments in which the outcome of any particular trial depends upon the outcome of the immediately preceding trial. In other words, the Markov chain is a sequence of 'states' through which a system passes at successive points in time.
9. There are two types of probability distributions, discrete and continuous probability distributions. In discrete probability distribution, the variable under consideration is allowed to take only a limited number of discrete values along with corresponding probabilities. On the other hand, in a continuous probability distribution, the variable under consideration is allowed to take on any value within a given range.
10. The probability function of binomial distribution is written as follows:

$$f(X = r) = {}^{n}C_{r} p^{r} q^{n-r}$$

$$r = 0, 1, 2, \ldots, n$$

where n = Numbers of trials

p = Probability of success in a single trial

$q = (1 - p)$ = Probability of failure in a single trial

r = Number of successes in n trials

11. The parameters of binomial distribution are p and n, where p specifies the probability of success in a single trial and n specifies the number of trials.
12. The use of binomial distribution is needed under the following circumstances:
    (a) When we have to find the probability of heads in 10 throws of a fair coin.
    (b) When we have to find the probability that 3 out of 10 items produced by a machine, which produces 8 per cent defective items on average, will be defective.
13. Poisson distribution is a discrete probability distribution that is frequently used in the context of operations research. Unlike binomial distribution,

Poisson distribution cannot be deduced on purely theoretical grounds based on the conditions of the experiment. In fact, it must be based on the experience, i.e. on the empirical results of past experiments relating to the problem under study.

14. Exponential distribution is used for describing a large class of phenomena, particularly in the area of reliability theory and in queuing models.
15. Normal distribution is the most important and frequently used continuous probability distribution among all the probability distributions. This is so because this distribution fits well in many types of problems. This distribution is of special significance in inferential statistics since it describes probabilistically the link between a statistic and a parameter.

4.7 QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the concept of probability.
2. What are the important theories of probability? Explain briefly.
3. Differentiate between objective and subjective probabilities.
4. How will you calculate the probability in the case of a simple event?
5. What is a mutually exclusive event?
6. What do you mean by compound events?
7. Define Jacob Bernoulli's theorem on probability.
8. What is Bayes' theorem? What is its importance in statistical calculations?
9. Describe Venn diagrams with reference to probability problems.
10. Define probability distribution and probability functions.
11. Describe binomial distribution and its measures.
12. How can a binomial distribution be fitted to given data?
13. Describe Poisson distribution and its important measures.
14. Poisson distribution can be an approximation of binomial distribution. Explain.
15. When is the Poisson distribution used?
16. Briefly describe exponential distribution.
17. Explain any six characteristics of normal distribution.
18. Write the formula for measuring the area under the curve.
19. Briefly explain the circumstances when the normal probability distribution can be used.

NOTES

Long-Answer Questions

1. What do you understand by the term probability? State and prove the addition theorem of probability.
2. Explain the meaning of mutually exclusive events, dependent and independent events giving suitable examples.
3. Distinguish between (a) *A priori* and empirical probabilities and (b) objective and subjective probabilities.
4. Define the concepts of sample space, sample points and events in context of probability theory. Also differentiate between simple and compound events.
5. (*a*) Explain the meaning of the following:
   - (*i*) Mathematical expectation
   - (*ii*) Bayes' theorem
   - (*iii*) Mutually exclusive events
   - (*iv*) Multiplication theorem of probability

   (*b*) Give your understanding of the following probability symbols:
   - (*i*) $P(A \cup B)$
   - (*ii*) $P(A / B)$
   - (*iii*) $P(A \cap B)$
6. A card is drawn at random from a pack of cards. What is the probability that it is either a 'heart' or the 'queen' of 'diamond'?
7. Four cards are drawn without replacement. Find the chance that they are all aces.
8. Three dice are thrown. What is the probability of at least one of the numbers turning up being greater than 4?
9. *A* speaks the truth in 80 per cent of the cases and *B* in 90 per cent of the cases. In what percentage of cases are they likely to contradict each other in stating the same fact?
10. If it is known from the past experience that rain falls at a station 12 days in every 30 days, then find the probability that in a given week four days will be wet and the remaining days dry.
11. In a group of equal numbers of men and women, 10 per cent men and 45 per cent women are unemployed. What is the chance that a person selected at random is employed?
12. *A* and *B* throw with one dice for a prize of Rs 11 which is to be won by the players who first throws 6. If *A* has the first throw, what are their respective expectations?

NOTES

13. Five men in a company of 20 are graduates. If 3 are picked out of 20 at random, what is the probability that they are all graduates? What is the probability of at least one is graduate?
14. Six coins are tossed simultaneously. What is the probability that they will fall with 4 heads and 2 tails up?
15. What is the probability of getting a number greater than 2 with an ordinary dice?
16. A pot contains 4 white and 6 red balls. Two drawings of 3 balls are made. Find the probability that the first drawing will give all the 3 balls white and the second all the 3 balls red, if the balls are replaced before the second trial.
17. There are 40 tickets in a lottery bearing numerals from 1 to 40. A ticket is drawn. What is the probability that it is multiple of 4 or 9?
18. If two six-sided dice, one marked *A* and the other *B* are thrown, what is the probability of getting one-dot side *A* dice and six-dots side of *B* dice?
19. In how many ways can the word EXAMINATION be arranged?
20. In an army battalion 3/5 of the soldiers are known to be married and the remainder 2/5 unmarried. Calculate the probability of getting 0, 1, 2, ..., 5 married soldiers in a row of 5 soldiers. If 500 rows each of 5 soldiers are standing on a ground, approximately, how many rows are expected to contain (a) all married soldiers and (b) all unmarried soldiers?
21. What is the chance that a leap year, selected at random, will contain 53 Saturdays?
22. It is found that 60 per cent of people between the ages 18 and 25 years pass their driving test on the first attempt. At the second attempt, the pass rate is 75 per cent of the first attempt pass rate. At the third attempt, the pass rate is only 1/2 that of the second attempt.

    What is the probability of anyone in that age range failing his or her test?
    (a) The first time
    (b) The second time
    (c) The third time
23. An egg importer finds that 15 per cent of the cartons of 6 eggs are damaged. A carton is picked at random, checked and returned to the consignment. This procedure is repeated a further three times. What is the probability that out of the 4 cartons so inspected, 3 were undamaged and 1 was damaged?
24. An iron monger's shop includes in its annual sale 21 factory-packed saucepans. It is known that 9 of them are red in colour, 8 are blue and 4 are yellow, although the colour tags have come off. A customer buys 3 at a bargain price. What is the probability that all 3 saucepans will be of the same colour?

NOTES

25. There are 10 white and 3 black balls in an urn. In a second urn, there are 3 white and 5 black balls. Two balls are transferred from the first urn to the second urn. Afterwards, one ball is drawn from the second urn. What is the probability that the ball so drawn is a white ball?
26. Three ships *A*, *B*, *C* sail from England to India. Chances in favour of their arriving safely are 2 : 5, 3 : 7 and 6 : 11 respectively. Find the chance that (a) they all arrive safely or (b) at least one arrives safely.
27. The odds against *A* solving a problem in statistics are 10 to 8 and the odds in favour of *B* solving the same problem are 12 to 9. What is the probability that if both of them attempt, the problem would be solved?
28. There are 4 roads joining two towns *A* and *B*, Mr *X* is allowed to go to town *B* from town *A* by one road and return by a different road. Find out the number of ways in which Mr *X* can perform the journey.
29. Forty lottery tickets numbered 1, 2, 3, ..., etc. are put in a bag. Two draws of one ticket each are made, the ticket after the first draw is replaced. What is the probability that in the first draw it is a multiple of 4 or 5 and, in the second, it is a multiple of 5 or 7?
30. Two cards are randomly drawn from a deck of 52 cards and thrown away. What is the probability of drawing an ace on a single draw from the remaining 50 cards?
31. Compare the probabilities of throwing 4 with one dice, 8 with two dice and 12 with three dice.
32. *A*, *B*, *C* in order cut a pack of cards, replacing them after each cut; find their respective chances of first cutting a heart.
33. Five coins whose faces are marked 2, 3 are thrown; what is the chance of obtaining a total of 12?
34. *A* makes a bet with *B* of 5 shilling to 2 shilling that in a single throw with two dice he will throw seven before *B* throws four. Each has a pair of dice and they throw simultaneously until one of them wins, equal throws being disregarded; find *B*'s expectations.
35. *A* and *B* throw alternately with a pair of dice. *A* wins if he throws 6 before *B* throws 7 and *B* wins if he throws 7 before *A* throws 6. If *A* begins, show that his chance of winning is 30/61.
36. If three squares are chosen at random on a chessboard, show that the chance that they should be in a diagonal line is 7/744.
37. *A* can hit a target four times in 5 shots; *B* three times in 4 shots; *C* twice in 3 shots. They fire a volley; what is the probability that two shot at least hit?
38. Three groups of children contain, respectively, 3 girls and 1 boy; 2 girls and 2 boys; 1 girl and 3 boys. One child is selected at random from each group. Show that the chance that the three selected consist of 1 girl and two boys is 13/32.

39. From each of 4 married couples, one of the married partners is selected at random. Show that the probability of their being all of one sex is 1/8.
40. There are three urns. Urn 1 contains 3 red and 7 green balls; urn 2 has 5 red and 3 green balls and urn 3 contains 8 red and 4 green balls. One red ball is drawn from one of the urns. What is the probability that it came from (a) urn 1 and (b) urn 3?
41. If it rains, an umbrella salesman can earn Rs 300 per day. If it is a clear sky, he can lose Rs 60 per day. What is his expectation if the probability of rain is 0.3?
42. An experiment succeeds twice as often as it fails. Find the chance that in the next six trials there will be at least four successes.
43. On average, out of 10 games a player wins 3 loses 5 and the others are drawn. Find the chance that out of the next 10 games he wins exactly 3 and loses exactly 5.
44. What are the odds against throwing seven twice at least in three throws with dice?
45. *A* and *B* play chess and *A* wins on average 2 games out of 3. Find the chance of *A* winning exactly 4 games out of the first 6, drawn games being disregarded.
46. Assume that a box contains 10 red balls and 6 black balls. Two balls are drawn, one at a time, without replacing the first ball. For this experiment, compute the following probabilities:
    (a) $P(B_2/B_1)$
    (b) $P(R_2/B_1)$
    (c) $P(R_2/R_1)$
47. Assume that there are two urns with the red and black balls as shown in the following:

6 *R*		8 *R*
4 *B*		2 *B*
No. I		No. II

    There is equal probability of choosing each urn. You take an urn, draw one ball and find that it is red. What is the probability that you drew the ball from urn 1? If the ball drawn happens to be black, what is the probability that it is from urn 2?
48. The manager of a certain business corporation in response to a question at lunch said that there was about one chance in a million that his corporation would fail as a result of a depression in the next 10 years. Later, he informed the shareholders of the corporation that even if there is a depression, the probability that his corporation would survive is 999/1000. But the corporation's monthly bulletin contains a signed article by the said manager wherein he says that the chances of a depression in the next 10 years are 1 in 100. Is the manager consistent in his probability assessments?

NOTES

49. (a) If A and B are independent, show that probability (A/B) = probability (A).
    (b) A family has two children. Assume that the probability of a boy is 1/2. Find the conditional probability of two boys given that one of the children is a boy.
    (c) A pair of dice have been thrown. Let A be the event that one dice is a 3 or 4. Let B be the event that other dice is a 4 or 5. Are the events A and B mutually exclusive?
50. In a certain school examination, results showed that 10 per cent of the students failed in mathematics, 12 per cent failed in English and 2 per cent failed in both mathematics and English. A student is selected at random from the school roll. Are the event 'Student failed in mathematics' and the event 'Student failed in English' independent?
51. (a) Explain the meaning of the Bernoulli process pointing out its main characteristics.
    (b) Give a few examples narrating some situations wherein binomial pr distribution can be used.
52. State the distinctive features of the binomial, Poisson and normal probability distributions. When does a binomial distribution tend to become a normal and a Poisson distribution? Explain.
53. Explain the circumstances when the following probability distributions are used:
    (a) Binomial distribution
    (b) Poisson distribution
    (c) Exponential distribution
    (d) Normal distribution
54. Certain articles have been produced of which 0.5 per cent are defective and the articles are packed in cartons each containing 130 articles. What proportion of cartons are free from defective articles? What proportion of cartons contain two or more defective articles?

    (Given $e^{-0.5}$=0.6065).
55. The following mistakes per page were observed in a book:

No. of Mistakes Per Page	*No. of Times the Mistake Occurred*
0	211
1	90
2	19
3	5
4	0
Total	345

Fit a Poisson distribution to the given data and test the goodness of fit.

56. In a distribution exactly normal, 7 per cent of the items are under 35 and 89 per cent are under 63. What are the mean and standard deviation of the distribution?
57. Assume the mean height of soldiers to be 68.22 inches with a variance of 10.8 inches. How many soldiers in a regiment of 1000 would you expect to be over six feet tall?
58. Fit a normal distribution to the following data:

Height in inches	*Frequency*
60–62	5
63–65	18
66–68	42
69–71	27
72–74	8

4.8 FURTHER READING

Anderson, David R., Dennis J. Sweeney and Thomas A. Williams. *Essentials of Statistics for Business and Economics*. Mumbai: Thomson Learning, 2007.

Anderson, David R., Dennis J. Sweeney and Thomas A. Williams. *Quantitative Methods for Business*. Mumbai: Thomson Learning, 2005.

Bhardwaj, R.S. *Business Statistics*. New Delhi: Excel Books, 2000.

Chandan, J.S. *Business Statistics.* New Delhi: Vikas Publishing House, 2004.

Gupta, C.B. and Vijay Gupta. *An Introduction to Statistical Methods*. New Delhi: Vikas Publishing House, 2004.

Hooda. R.P. *Statistics for Business & Economics.* New Delhi: Macmillan India Ltd., 2004.

Kothari C.R. *Quantitative Techniques*. New Delhi: Vikas Publishing House, 1984.

Levin, Richard I. and David S. Rubin. *Statistics for Business*. New Delhi: Prentice Hall of India, 1990.

Monga, G.S. *Mathematics and Statistics for Economics*. New Delhi: Vikas Publishing House.

Sancheti D.C. and V.K. Kapoor. *Business Mathematics*. New Delhi: Sultan Chand & Sons.

Zameeruddin Qazi, V.K. Sharma and S.K. Bhambri. *Business Mathematics*. New Delhi: Vikas Publishing House, 2008.

APPENDIX: STATISTICAL TABLES

(I) Some binomial distributions: $P(r \leq r_0/n, p)$

(II) Values of $e^{-\lambda}$ (for calculating Poisson probabilities)

(III) Area under standard normal distribution

(IV) Cumulative normal distribution

(V) Some critical values of '*t*'

NOTES

APPENDIX

NOTES

TABLE I

Some Binomial Distributions: $P(r \leqslant r_0/n, p)$

n	r_0	p 0.10	p 0.25	p 0.40	p 0.50
1	0	0.9000	0.7500	0.6000	0.5000
	1	1.0000	1.0000	1.0000	1.0000
5	0	0.5905	0.2373	0.0778	0.0313
	1	0.9185	0.6328	0.3370	0.1875
	2	0.9914	0.8965	0.6826	0.5000
	3	0.9995	0.9844	0.9130	0.8125
	4	0.9999	0.9990	0.9898	0.9687
	5	1.0000	1.0000	1.0000	1.0000
10	0	0.3487	0.0563	0.0060	0.0010
	1	0.7361	0.2440	0.0463	0.0108
	2	0.9298	0.5256	0.1672	0.0547
	3	0.9872	0.7759	0.3822	0.1719
	4	0.9984	0.9219	0.6330	0.3770
	5	0.9999	0.9803	0.8337	0.6230
	6	1.0000	0.9965	0.9452	0.8281
	7	1.0000	0.9996	0.9877	0.9453
	8	1.0000	1.0000	0.9983	0.9892
	9	1.0000	1.0000	0.9999	0.9990
	10	1.0000	1.0000	1.0000	1.0000
20	0	0.1216	0.0032	0.0000	0.0000
	1	0.3917	0.0243	0.0005	0.0000
	2	0.6768	0.0912	0.0036	0.0002
	3	0.8669	0.2251	0.0159	0.0013
	4	0.9567	0.4148	0.0509	0.0059
	5	0.9886	0.6171	0.1255	0.0207
	6	0.9975	0.7857	0.2499	0.0577
	7	0.9995	0.8981	0.4158	0.1316
	8	0.9999	0.9590	0.5955	0.2517
	9	1.0000	0.9861	0.7552	0.4119
	10	1.0000	0.9960	0.8723	0.5881
	1 1	1.0000	0.9990	0.9433	0.7483
	12	1.0000	0.9998	0.9788	0.8684
	13	1.0000	1.0000	0.9934	0.9423
	14	1.0000	1.0000	0.9983	0.9793
	15	1.0000	1.0000	0.9996	0.9941
	16	1.0000	1.0000	1.0000	0.9987
	17	1.0000	1.0000	1.0000	0.9998
	18	1.0000	1.0000	1.0000	1.0000
	19	1.0000	1.0000	1.0000	1.0000
	20	1.0000	1.0000	1.0000	1.0000

TABLE II

Values of $e^{-\lambda}$ (For Computing Poission Probabilities)

λ *(Mean)*	$e^{-\lambda}$	λ *(Mean)*	$e^{-\lambda}$	λ *(Mean)*	$e^{-\lambda}$	λ *(Mean)*	$e^{-\lambda}$
0.1	0.90484	2.6	0.07427	5.1	0.00610	7.6	0.00050
0.2	0.81873	2.7	0.06721	5.2	0.00552	7.7	0.00045
0.3	0.74042	2.8	0.06081	5.3	0.00499	7.8	0.00041
0.4	0.67032	2.9	0.05502	5.4	0.00452	7.9	0.00037
0.5	0.60653	3.0	0.04979	5.5	0.00409	8.0	0.00034
0.6	0.54881	3.1	0.04505	5.6	0.00370	8.1	0.00030
0.7	0.49659	3.2	0.04076	5.7	0.00335	8.2	0.00027
0.8	0.44933	3.3	0.03688	5.8	0.00303	8.3	0.00025
0.9	0.40657	3.4	0.03337	5.9	0.00274	8.4	0.00022
1.0	0.36788	3.5	0.03020	6.0	0.00248	8.5	0.00020
1.1	0.33287	3.6	0.02732	6.1	0.00224	8.6	0.03018
1.2	0.30119	3.7	0.02472	6.2	0.00203	8.7	0.00017
1.3	0.27253	3.8	0.02237	6.3	0.00184	8.8	0.00015
1.4	0.24660	3.9	0.02024	6.4	0.00166	8.9	0.00014
1.5	0.22313	4.0	0.01832	6.5	0.00150	9.0	0.00012
1.6	0.20190	4.1	0.01657	6.6	0.00136	9.1	0.00011
1.7	0.18268	4.2	0.01500	6.7	0.00123	9.2	0.00010
1.8	0.16530	4.3	0.01357	6.8	0.00111	9.3	0.00009
1.9	0.14957	4.4	0.01228	6.9	0.00101	9.4	0.00008
2.0	0.13534	4.5	0.01111	7.0	0.00091	9.5	0.00007
2.1	0.12246	4.6	0.01005	7.1	0.00083	9.6	0.00007
2.2	0.1 1080	4.7	0.00910	7.2	0.00075	9.7	0.00006
2.3	0.10026	4.8	0.00823	7.3	0.00068	9.8	0.00006
2.4	0.09072	4.9	0.00745	7.4	0.00061	9.9	0.00005
2.5	0.08208	5.0	0.00674	7.5	0.00055	10.0	0.00005

NOTES

NOTES

TABLE III

Area under Standard Normal Distribution between the Mean and Successive Value of *Z*

Z	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4804	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

TABLE IV
Cumulative Normal Distribution
(Extract from the Table Concerning the Area under the Normal Curve)

Z	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8454	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8927	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.99406
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96851	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98125	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98773	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.990097	0.990358	0.090613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613

Note: When $Z = 4.09$, the table value is 0.99997843 (i.e. equal to 1 for all practical purposes).
(This table contains the area of left half of the curve (i.e. 0.5) plus the relevant area of the right half for different values of Z)

NOTES

NOTES

TABLE V
Some Critical Values of '*t*'

Degrees of Freedom	*Level of Significance*		
	1%	*5%*	*10%*
1	63.657	12.706	6.314
2	9.925	4.303	2.920
3	5.841	3.182	2.353
4	4.604	2.776	2.132
5	4.032	2.571	2.015
6	3.707	2.447	1.943
7	3.499	2.365	1.895
8	3.355	2.306	1.860
9	3.250	2.262	1.833
10	3.169	2.228	1.812
11	3.106	2.201	1.796
12	3.055	2.179	1.782
13	3.012	2.160	1.771
14	2.977	2.145	1.761
15	2.947	2.131	1.753
16	2.921	2.120	1.746
17	2.898	2.110	1.740
18	2.878	2.101	1.734
19	2.861	2.093	1.729
20	2.845	2.086	1.725
21	2.831	2.080	1.721
22	2.819	2.074	1.717
23	2.807	2.069	1.714
24	2.797	2.064	1.711
25	2.787	2.060	1.708
26	2.779	2.056	1.706
27	2.771	2.052	1.703
28	2.763	2.048	1.701
29	2.756	2.045	1.699
α	2.576	1.960	1.645

Note: These table values of '*t*' are in respect of two-tailed tests. If we use the *t*-distribution for one-tailed test then we are interested in determining the area located in one tail. So to find the appropriate *t*-value for a one-tailed test say at a 5% level with 12 degrees of freedom, then we should look in the above table under the 10% column opposite the 12 degrees of freedom row. (This value will be 1.782). This is true because the 10% column represents 10% of the area under the curve contained in both tails combined, and so it also represents 5% of the area under the curve contained in each of the tails separately.