Quantitative Methods

The Icfai University Press

ISBN : 81-7881-929-5

Ref. No. QM – 052008UG041

For any clarification regarding this book, please write to FedUni giving the reference number of the book, and the page number.

While every possible care has been taken in preparing the book, FedUni welcomes suggestions from students for improvement in future editions.

Quantitative Methods

Contents

Quantitative Methods

Detailed Curriculum

Basics of Mathematics: *Number System*: Real Numbers – Inequalities and Intervals – Least Common Multiple (LCM) – Highest Common Factor (HCF) – Basic Operations on Fractions.

Polynomials: Addition and Subtraction of Like Terms and Unlike Terms – Multiplication and Division of Like Terms and Unlike Terms – Dimensions and Degree of an Expression.

Simultaneous and Quadratic Equations: Solving Simultaneous Equations – Solving Quadratic Equations.

Indices: Theory of Indices – Laws of Indices.

Progressions: Arithmetic Progression – Geometric Progression – Harmonic Progression.

Permutations and Combinations: Permutations – Combinations.

Logarithms: Rules of Logarithms – Transforming the Base of Logarithms.

Introduction to Statistics: Origin and Growth of Statistics – Applications of Statistics – Collection of Data – Significance of Computers in Statistics.

Sampling: Census and Sample Method – Theoretical Basis of Sampling – Methods of Sampling – Size of Sample – Merits and Limitations of Sampling – Sampling and Non-sampling Errors.

Classification and Tabulation of Data: Meaning and Objectives of Classification – Types of Classification – Tabulation of Data – Parts of Table – Rules of Tabulation – Types of Tables – Table Review – Frequency Distributions.

Diagrammatic and Graphic Presentation: Significance of Diagrams and Graphs – Diagrams – Types of Diagrams – Graphs – Types of Graphs – Techniques of Constructing Graphs – Difference between Diagrams and Graphs – Limitations of Diagrams and Graphs.

Measures of Central Tendency: Meaning and Objectives of Averaging – Types of Averages – Appropriateness of the Three Principal Averages – Relationship among the Mathematical Averages – Choice of a Suitable Average – Additional Illustrations.

Measures of Dispersion: Meaning of Dispersion – Properties of a Good Measure of Variation/Dispersion – Significance of Measuring Variation/Dispersion – Methods of Studying Variation/Dispersion – Relationship between Quartile Deviation, Standard Deviation, and Mean Deviation – Graphical Method – Lorenz Curve – Additional Illustrations.

Skewness: Types of Distributions – Meaning of Skewed Distribution – Measures of Skewness – Additional Illustrations.

Correlation: Meaning and Definition of Correlation – Types of Correlation – Significance of Correlation – Methods of Studying Correlation – Karl Pearson's Coefficient of Correlation – Coefficient of Correlation and Probable Error – Rank Correlation Coefficient – Coefficient of Determination – Concurrent Deviation Method – Additional Illustrations.

Regression Analysis: Meaning and Definition of Regression Analysis – Application of Regression Analysis – Difference between Correlation and Regression Analysis – Regression Line – Regression Equations – Standard Error of Estimate – Additional Illustrations.

Index Numbers:

The Concept of Index Numbers – Types of Index Numbers – Methods of Constructing Index Numbers – Aggregates Method – Average of Relatives Method – Value Index Numbers – Tests for Consistency – Consumer Price Index Number – Additional Illustrations.

Time Series Analysis: Time Series Analysis – Procedure for Fitting a Straight Line – Components of Time Series Analysis (Secular Trend, Cyclical Variation, Seasonal Variation (with ratio of moving averages) and Irregular Variation) – Comprehensive Illustration – Time Series Analysis in Forecasting – Additional Illustrations.

Probability: The Concept of Probability – Approaches to Probability – Probability Rules – Bayes' Theorem – Additional Illustrations.

Theoretical Distributions: Random Variable – Expected Value – Theoretical Distributions – Binomial Distribution – Poisson Distribution – Normal Distribution – Additional Illustrations.

Linear Programming: Meaning of Linear Programming – Review of Linear Functions – The Graphical Method of Linear Programming – The Simplex Method of Linear Programming – Post Optimal Analysis – Duality – Additional Illustrations.

Chapter I

Basics of Mathematics

Lesson 1

Number System

After reading this lesson, you will be conversant with:

- Real Numbers
- Inequalities and Intervals
- Least Common Multiple (LCM)
- Highest Common Factor (HCF)
- Basic Operations on Fractions

Introduction

We know that one has to deal with numbers in day-to-day life, irrespective of one's inclination and field of work. One cannot also refute the fact that in this age there is nothing like learning mathematics for its own sake or for that matter studying psychology for its own sake. One should understand the fact that the emphasis today is on application, better decision-making and in trying to decipher what the future may look like. This scenario calls for professionals working in diverse fields to possess multiple skill sets. In most of the cases, it calls for understanding and application of elementary principles of other disciplines. To be more clear, a psychologist may not be as effective as he can be, if he misses out on, say, basic computer skills or a geneticist cannot exceed in career if he does not possess basic knowledge in statistics, or for that matter a student cannot pursue a career in finance without knowing the basics of mathematics. Keeping this in view, we present below seven topics in the basics of mathematics. The first topic 'Real Numbers' deals with number system and other basics like finding the Least Common Multiple and Highest Common Factor of the quantities given. The second topic 'Polynomials' will brush up your knowledge and provide insights as to how different quantities interact with each other. The third topic 'Simultaneous equations' will give an overview of how to solve simultaneous equations by different methods and also a method which can be applied only to quadratic equations. The fourth topic 'Theory of Indices' will demystify how one would obtain x^9 as a result of multiplying x^6 and x^3. The fifth topic 'Progressions' will give you a better perspective of what series is. The sixth topic 'Permutations and Combinations' will prepare you to deal more confidently with selections and arrangements of things and the last topic covered is 'Logarithms'. All these topics, we hope, will help you understand the beauty and subtlety of the science called 'Mathematics' at least to some extent.

x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x

In this part, we will first look at how the number system has developed over a period of time and then at some of the important algebraic properties of real numbers.

Natural Numbers: To begin with, we have counting numbers. These numbers are also known as Natural numbers and are denoted by a symbol 'N'. These numbers are obtained by adding one to the previous number. In other words, once we know the first element we can obtain the elements following it by adding 1 to the successive elements. That is, we can have infinite (innumerable or a large number) number of them. Since natural numbers are infinite, we find it convenient to express them as a set. By set we mean a collection of well-defined objects. Each individual object is also referred to as an element of that particular set. We should be clear that the concept of set can be used to represent infinite as well as finite number of elements. A simple example for finite number of elements would be the vowels a, e, i, o and u expressed as a set. Generally, the set of natural numbers is represented as:

$$N = \{1, 2, 3, ...\}.$$

As we look at other sets we will realize a number of drawbacks that the set of natural numbers suffer from.

Whole Numbers: Observe that natural numbers did not have a zero. This shortcoming is made good when we consider the set of whole numbers. The set of whole numbers is denoted by 'W' and is expressed as:

$$W = \{0, 1, 2, 3, ...\}$$

Integers: The set of whole numbers also does not satisfy all our requirements as on observation, we find that it does not include negative numbers like –2, –7 and so on. To overcome this shortcoming, a set of elements called Integers had been constituted. This set is denoted by the symbol 'I'. It is represented as:

$$I = \{.....-3, -2, -1, 0, 1, 2, 3, ...\}$$

This set is definitely superior to natural and whole numbers in the sense that it caters to a larger audience as compared to the other number systems seen above.

Rational Numbers: Although the set of integers caters to a larger audience, it is inadequate. This inadequacy has led to the formulation of Rational numbers. Rational numbers are of the form $\frac{p}{q}$, where p and q are integers and $q \neq 0$. The numbers like $\frac{2}{3}, \frac{-5}{4}$ are examples of rational numbers. The set of rational numbers is denoted by Q and generally expressed as:

$$Q = \{....., \frac{-2}{5}, \frac{-1}{4}, \frac{0}{1}, \frac{3}{5}, \frac{6}{7}, \frac{7}{8}, ……\}$$

In the set of rational numbers, if you consider any of the elements say –2/5, we observe that the quotient is – 0.4. Similarly if you consider 7/8, the quotient is 0.875. In both these cases, the decimal part is terminating. By terminating, we understand that the division process is coming to an end. Now, in the same set, consider the element 6/7. For this number, the quotient is 0.857142857142..... In this case, we observe that the decimal part (i) is not terminating, and (ii) it is repeating.

But occasionally we also find decimals which neither terminate nor repeat. For instance, consider a number like 65/67. The quotient is of the form 0.970149253..... In this quotient we neither find the decimal terminating nor repeating. Numbers whose decimals are non-terminating and non-repeating are included in a set of numbers called irrational numbers.

REAL NUMBERS

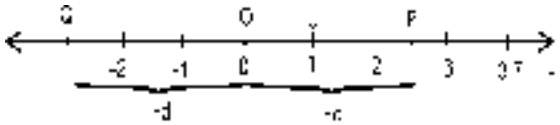
All the number sets we have seen above put together comprise the real numbers. Real numbers are also inadequate in the sense that they do not include a quantity which is the square root of a negative number. One has to look towards complex numbers to deal with such quantities. Since this is beyond the scope of this book, here we discuss only real numbers.

Defining Real Numbers

The numbers used to measure quantities such as length, area, volume, body temperature, GNP, growth rate etc., are called real numbers.

Another definition of real numbers uses a concept called number line (we will learn about it shortly). According to this definition, a real number is any number that is the co-ordinate of the point on a real number line.

Sets of real numbers and relations among such sets can often be visualized by the use of a number line or co-ordinate axis. A number line is constructed by fixing a point O called the origin and another point U called the unit point on a horizontal straight line L. It is shown below. Therefore, on a number line we observe that positive numbers are in the increasing order as we move to the right of the origin whereas the negative numbers are in the decreasing order as we move to the left of the origin.



Each point P on the line L is now assigned a “numerical address” or co-ordinate “x” representing its signed distance from the origin, measured in terms of the given unit. Thus $x = \pm d$, where d is the distance between O and P (shown above). The plus sign or minus sign is used to indicate whether P is to the right of the origin or to the left of the origin. If the resulting number line is drawn to some

scale, each point P has a corresponding numerical value and to each real numerical value "x" there will be corresponding unique point P on the number line. This can be observed on the number line. In other words for a point say 3.7, there is only a single point on the number line and for a point Q there will be unique numerical value.

Basic Algebraic Properties of Real Numbers

These can be expressed in terms of two fundamental operations of addition and multiplication.

If a, b and c are any three real numbers, then;

1. i. $a + b = b + a$

   This property is called commutative property of addition. According to this property, addition can be carried out in any order irrespective of which we obtain the same result.

   ii. $a.b = b.a$

   This property is called commutative property of multiplication.

2. i. $(a + b) + c = a + (b + c)$

   This property is referred to as associative property of addition. According to this property, elements can be grouped according to any manner and irrespective of the grouping, we obtain the same result.

   ii. $(a.b).c = a.(b.c)$

   This property is referred to as the associative property of multiplication.

3. $a.(b + c) = a.b + a.c$ or $(a + b).c = a.c + b.c$

   This property is referred to as distributive property. This is generally employed to expand a product into a sum or the other way round. That is, to rewrite a sum as a product.

4. i. $a + 0 = 0 + a = a$

   This property is referred to as identity property under addition. That is, when 0 is added to a real number, we get back the number itself. Thus 0 is the identity element under addition.

   ii. $a.1 = 1.a = a$

   This property is referred to as identity property under multiplication. That is, when a real number is multiplied by 1, we get back the same number.

   Thus the element 1 is the multiplicative identity.

5. i. $a + (–a) = (–a) + a = 0$

   This property is referred to as inverse property under addition. According to this property, for every element a, there exists another element – a such that the addition of the both gives us zero. The element – a is referred to as the additive inverse of the element a. On a number line, an element and its additive inverse lie at equi-distant from the origin.

   ii. $a \times \frac{1}{a} = \frac{1}{a} \times a = 1$

   This property is referred to as inverse property under multiplication. According to this property, for every element a, $a \neq 0$, there exists another element 1/a such that the multiplication of a and 1/a results in 1. The element 1/a is referred to as multiplicative inverse element.

6. i. If $a + x = a + y$, then $x = y$.

   This property is referred to as the cancellation property. According to this property a constant quantity when present on both sides of the equation can be cancelled without disturbing the balance which exists between the expressions.

ii. If $a \neq 0$ and $ax = ay$, then $x = y$.

This property is referred to as the cancellation property under multiplication.

7. i. $a.0 = 0.a = 0$

This property is referred to as the zero factor property. According to this property, any real number a, if multiplied by zero would yield a zero. This can be also put as: if one of the factors happens to be zero, irrespective of other factors, the product of all these factors would yield a zero.

ii. If $a.b = 0$, then $a = 0$ or $b = 0$ or both.

According to this property, the product of any two real numbers a and b is zero, if one of them happens to be zero, that is either $a = 0$ or $b = 0$ or both of them happen to be equal to zero.

INEQUALITIES AND INTERVALS

In this part, we look at what inequalities and intervals are. If it is given that a real number 'p' is not less than another real number 'q', we understand that either p should be equal to q or p should be greater than q. We express the same as $p = q$ or $p > q$ or $p \geq q$. If p is greater than q, then, is not $p - q$ a positive number? It is. Statements like $p < q$ or $p > q$ are called inequalities. We employ the concept of inequalities to understand sets referred to as intervals. We define an interval as a range of values from which the real number "p" is likely to assume values. In this context, interval is more restricted as compared to a set of real numbers which ranges from $+\infty$ to $-\infty$. We should note that '∞' is not a real number and it is employed only because of convenience.

Rules for Inequalities

Here we look at only the rules without going into their proofs. They are:

1. $a < b$, if, and only if $b - a > 0$.
2. If $a < b$ and $b < c$, then $a < c$.
3. If $a < b$, then $(a + c) < (b + c)$.
4. If $a < b$, then $-a > -b$.
5. If $a < b$ and
   i. $x = 0$, then $ax = bx$
   ii. $x < 0$, then $ax > bx$
   iii. $x > 0$, then $ax < bx$
6. If $0 < a < b$, then $0 < 1/b < 1/a$.

Bounded Intervals

Let a and b be fixed real numbers such that $a < b$ on a number line. The different types of intervals we have are:

i. **The open interval (a, b):** We define an open interval (a, b) with end points a and b as a set of all real numbers "x", such that $a < x < b$. That is, the real number x will be taking all the values between a and b. An important point to consider in this case is the type of brackets used. Generally open intervals are denoted by ordinary brackets ( ).

ii. **The closed interval [a, b]:** We define a closed interval [a, b] with end points a and b as a set of all real numbers "x", such that $a \leq x \leq b$. In this case the real number x will be taking all the values between a and b inclusive of the end points a and b. Generally closed intervals are denoted by [ ] brackets.

iii. **The half open interval [a, b):** We define a half open interval [a, b) with end points a and b as a set of all real numbers "x", such that $a \leq x < b$. In this case, the real number x will be taking all the values between a and b, inclusive of only a but not b.

iv. **The half open interval (a, b]:** We define a half open interval (a, b] with end points a and b as a set of all real numbers "x", such that $a < x \leq b$. In this case the real number x will be taking all the values between a and b, inclusive of only b but not a.

Unbounded Intervals

Intervals which extend indefinitely in both the directions are known as unbounded intervals. These are written with the aid of symbols $+\infty$ and $-\infty$. The various types of intervals, if "a" happens to be a real number, are:

i. $(a, +\infty)$ is the set of all real numbers x such that $a < x$.

ii. $(-\infty, a)$ is the set of all real numbers x such that $x < a$.

iii. $[a, +\infty)$ is the set of all real numbers x such that $a \leq x$.

iv. $(-\infty, a]$ is the set of all real numbers x such that $x \leq a$.

Absolute Value of a Number

At times we consider only the magnitude of the number without attaching much importance to its direction. Under these circumstances, the sign attached with the number becomes redundant. This concept can also be expressed in terms of the absolute value of a number. The absolute value is nothing but the magnitude of the real number. In this context, the absolute value can never be negative as distance cannot be negative. That is $|-3| = |+3| = 3$.

The same concept when considered in the light of real numbers will be

$|-x| = x$, if $x < 0$

$|x| = x$, if $x > 0$.

Apart from these sets, we also have even numbers, odd numbers, prime and composite numbers. A number is defined as an even number if it is exactly divisible by 2; otherwise, it is an odd number. A number is said to be a prime number if it has its factors as 1 and itself. That is, if we observe the factors of numbers like 2, 3, 5, 7, 11 etc., we find that they do not have other factors other than 1 and themselves. Numbers like 4, 15 and 96 which have factors other than 1 and themselves are called composite numbers.

LEAST COMMON MULTIPLE (LCM)

Before we look at this, let us learn what a multiple is. Take any number say 3. Multiply this number with natural numbers. We obtain 3, 6, 9, 12, 15, 18,..... The resulting numbers, that is 3, 6, 9, 12, etc., are called multiples of 3. Now to understand the concept of least common multiple consider two numbers 4 and 6. The multiples of 4 and 6 are:

Multiples of 4 = 4, 8, 12, 16, 20, 24, 28, 32, 36, ...

Multiples of 6 = 6, 12, 18, 24, 30, 36, 42, 48, 54, ...

From this we can observe that the common multiples to both 4 and 6 are 12, 24, 36 etc. Of these common multiples the least is 12. This gives our least common multiple as 12. Therefore the least common multiple or LCM is defined as that quantity which is divisible by the quantities of which it is an LCM without a remainder. In our example, 12 is the least possible quantity that can be divided by 4 and 6 without leaving any remainder.

Now, let us take another example and check whether the above method of finding the LCM can be applied to all the problems. Find the LCM of 3, 5 and 7. We list the multiples of these three numbers.

Multiples of 3 = 3, 6, 9, 12, 15, 18, 21, ...

Multiples of 5 = 5, 10, 15, 20, 25, 30, 35, ...

Multiples of 7 = 7, 14, 21, 28, 35, 42, 49, ...

We find that this method takes substantial amount of time. We look at another method given below. This method depends on the premise that most of the numbers can be expressed in terms of prime numbers. That is, if you are given to compute the LCM of numbers 35 and 42, the first thing we do is to express them in terms of prime numbers. That will be

$35 = 5 \times 7$

$42 = 2 \times 3 \times 7$

Since we ought to get a number which is the least common multiple of both these numbers, we choose our prime factors in a manner that the resultant number includes the factors for at least the minimum of times it is present in either of given numbers. Therefore from the prime factors above we choose 5. Observe that 5 is a factor of 35 and not 42. As it ought to be present at least once we consider it. Then we choose 7. Both the numbers have it once, therefore we take it once; the number 2 is a factor of 42 and not of 35. By the above logic we choose it once. The same with 3 also. Therefore, the product of these numbers gives us $2 \times 3 \times 5 \times 7 = 210$ which is our required LCM. Consider another example. Find the LCM of 24 and 45. We express them in terms of prime factors.

$24 = 2 \times 2 \times 2 \times 3$

$45 = 3 \times 3 \times 5$

From these we know that 2 should be present at least thrice, 3 should be present at least twice and 5 at least once. All these factors multiplied will give us our LCM. It will be $2 \times 2 \times 2 \times 3 \times 3 \times 5 = 360$.

When bigger numbers are given even the above method can prove tedious.

We illustrate another method, which is perhaps more easier than what we have seen above. Compute the LCM of 28, 54 and 81. In this method, we write these three numbers as shown below. We start dividing these by an appropriate prime number, which in this case happens to be 2. The quotients are put as shown below. Then we start dividing these quotients by a prime number, this prime number may be the same one which has been already employed or it can be a different prime number (remember that we do not have any rule which specifies that a particular number should be used for so many times). This process is repeated until we have prime numbers left in the quotient part which also indicates that we cannot divide these numbers any further. This is shown below.

2	28, 54, 81
3	14, 27, 81
3	14, 9, 27
3	14, 3, 9
2	14, 1, 3
	7, 1, 3

We observe that the quotient part has 7, 1 and 3 which are prime numbers. Now we multiply all the prime numbers with which we divided the given numbers and those which are in the last row of the quotient part. That is, $2 \times 3 \times 3 \times 3 \times 2 \times 7 \times 1 \times 3 = 2268$. Thus our required LCM is 2268.

HIGHEST COMMON FACTOR (HCF)

We know that a factor is a quantity which divides the given quantity without leaving any remainder. Similar to LCM above we can find a Highest Common Factor (HCF) of the given numbers. Let us look at its definition first. The highest common factor is a quantity obtained from the given quantities and which divides each of them without leaving a remainder. We understand this by taking an example.

Example 1

Find the HCF of 49 and 63.

The factors of 49 are 1, 7 and itself. The factors of 63 are 1, 3, 7, 9, 21 and itself. The common factors are 1 and 7. The highest of these is 7, which is the HCF we require.

This is one of the methods to obtain the HCF. This method may prove tedious if we are given bigger numbers and more of them. When such quantities are given, we follow division method as shown below (this method is shown for numbers in the above example).

In this method, the first step constitutes dividing the larger quantity by the smaller quantity and subtract it as shown to obtain a remainder (it is not necessary that we ought to get a remainder in all the cases). Then the divisor, 49 (in our case, 49 was the divisor and 63 the dividend, 1 the quotient and 14, the remainder) becomes the dividend and the remainder (14) which we obtained earlier becomes the divisor. We continue doing this until the remainder is 0 as shown below. The last divisor is our HCF.

```
49) 63 (1
    49
  ---------
  14) 49 (3
     42
    --------
     7) 14 (2
        14
       -----
         0
```

That is, 7 is the HCF of the numbers 49 and 63.

Now let us consider three quantities and obtain the HCF for them.

Example 2

Find the Highest Common Factor of 54, 72 and 150.

First we consider 54 and 72. The HCF for these two quantities is calculated as follows:

```
54) 72 (1
    54
  ---------
  18) 54 (3
     54
    -------
       0
    -------
```

The HCF is 18. Now we consider 18 and 150 and obtain the HCF for these two quantities. It will be obtained as follows.

```
18) 150 (8
    144
  ---------
     6) 18 (3
        18
       -------
          0
       -------
```

We observe that 6 is the highest common factor for these two quantities. That is, 6 is the HCF of the three given quantities.

(**Note:** The concept of LCM and HCF can be applied to expressions as well.)

BASIC OPERATIONS ON FRACTIONS

A simple example of fraction would be a rational number of the form $\frac{p}{q}$, where $q \neq 0$. In fractions also we come across different types of them. The two fractions 3/4 and 1/4 are like fractions and the fractions 2/5 and 6/7 are unlike fractions. That is, fractions whose denominators are same are referred to as like fractions and the fractions like 2/5 and 6/7 are called unlike fractions as their denominators differ. Further when the numerator in a fraction is lower than the denominator, that fraction is referred to as proper fraction and the fraction in which the numerator is greater than the denominator, is referred to as improper fraction. Also a fraction like $3\frac{2}{5}$ is referred to as mixed fraction as it consists of an integer 3 and a fractional part 2/5.

Addition of Like Fractions: While adding like fractions the denominator will have the same term as that present in the individual quantities, while the numerator will be the sum of numerators present in the individual fractions.

We take an example.

Example 3

Add 2/5 and 7/5.

We have $\frac{2}{5}+\frac{7}{5}=\frac{2+7}{5}=\frac{9}{5}$

Subtraction of Like Fractions: This will be similar to addition of fractions. Only that the plus symbol should be replaced by the minus symbol. The subtraction operation for the above fractions will be

$$\frac{2}{5}-\frac{7}{5}=\frac{2-7}{5}=\frac{-5}{5}$$

Multiplication of Fractions: The multiplication of fractions will be much simpler. We multiply the numerators and the denominators respectively and express the product as a fraction. For the fractions 2/5 and 7/5, the product will be $\frac{2}{5}\text{ x }\frac{7}{5}=\frac{14}{25}$

Division of Like Fractions: If we have to divide one fraction with the other, we multiply the first one with the reciprocal of the second. For the fractions 2/5 and 7/5, the quotient will be:

$$\frac{2}{5}\text{ x }\frac{5}{7}=\frac{2}{7}$$

Addition of Unlike Fractions: This can be better understood with the help of an example only. Add 2/5 and 7/3. We begin by taking the LCM of the terms present in the denominators of the given fractions. In our case the LCM will be 5 x 3 = 15. We write that as shown below.

$$\overline{15}$$

Now we divide the LCM by the denominator of the first fraction. We obtain 15/5 = 3. In the numerator, the product of this term (3) and the term in the numerator of the first fraction (2), that is 2 x 3 = 6 is stated. It is shown below.

$$\frac{6+}{15}$$

We repeat the same procedure for the second fraction also. On division we obtain 15/3 = 5. Then we multiply 5 with the term in the numerator of the second term. We obtain 5 x 7 = 35 and write this term as shown below. The sum of these two terms gives us our required result.

$$\frac{6+35}{15}=\frac{41}{15}$$

Subtraction of Unlike Fractions: This is identical to what we have seen above except that the symbol has to be replaced. In our case it will be

$$\frac{6-35}{15}=\frac{-29}{15}$$

Multiplication of Unlike Fractions: This will be similar to multiplication of like terms we have seen before. For the fractions, 2/5 and 7/3, the product will be

$$\frac{2}{5}\text{ x }\frac{7}{3}=\frac{14}{15}$$

Division of Unlike Fractions: This will be similar to what we have seen in like terms. The quotient of the fractions 2/5 and 7/3 will be

$$\frac{2}{5}\text{ x }\frac{3}{7}=\frac{6}{35}$$

Reducing the Fractions to Lowest Terms: By 'reducing a fraction to its lowest terms' we understand that the numerator and the denominator of the fraction being reduced to lowest terms by dividing the numerator and the denominator by the same term. This we do repeatedly until it becomes clear that we cannot do it any further. This should be clear if we look at an example.

Example 4

Reduce 24/36 to its lowest terms. $\frac{24}{36}=\frac{12}{18}=\frac{6}{9}=\frac{2}{3}$. In the first step, we divide the numerator and the denominator by 2. The fraction gets reduced to 12/18. We divide this fraction again by 2. It stands reduced to 6/9. This we divide by 3. We obtain 2/3. Thus, we say that the fraction $\frac{24}{36}$ has been reduced to its lowest terms, which happens to be a fraction $\frac{2}{3}$.

SUMMARY

- There are different number sets like Natural numbers, Whole numbers, Integers and Rational numbers and all these sets are collectively called "Real numbers".
- The study also provides an insight into inequalities and intervals, LCM and HCF, and fractions.

Lesson 2

Polynomials

After reading this lesson, you will be conversant with:

- Addition and Subtraction of Like Terms and Unlike Terms
- Multiplication and Division of Like Terms and Unlike Terms
- Dimensions and Degree of an Expression

In arithmetic, we deal with numbers. In contrast to this, in algebra, we deal with symbols. These symbols are often represented by lower case alphabets. One of the advantages of using alphabets is that they are easily identifiable and make some sense even to a lay person. As against figures which stand for a definite value under all the circumstances, symbols in algebra are situation specific. That is, x in a certain problem may take 2 as its value while in some other problem it may take 3 or 11 or 154 as its value, while numbers remain the same whenever and wherever they are used. Another important facet of algebra is that we can operate with symbols themselves without assigning them any value whatsoever. The basic signs (+, –, x, ÷) or operators as they are otherwise known as, have the same meaning they hold in arithmetic. In algebra, we can have terms like 'a' or '7q' or '7p + b' or 'a + 8b – c'. If we consider terms like a + 8b – c or 7p + b, we find that they are a collection of symbols separated by signs and they may contain either one or two or more such symbols. All such collection of terms separated by signs are generally referred to as algebraic expressions. And as you can see, expressions may be simple or compound. A compound expression can be a binomial, trinomial or a polynomial expression. An expression like 'a' or '7q' is referred to as simple expression whereas expressions like 7p + b is referred to as binomial expression as it contains two terms and the expression a + 8b – c is referred to as a trinomial (it is a collection of three terms). Also expressions containing more than three terms are referred to as multinomials or polynomials.

When two or more quantities are multiplied with each other, the resultant number is referred to as the product of those quantities. The point to observe in this case is that while in arithmetic the product of two numbers is shown as 2 x 3, the same in algebra can be shown as ab or a x b or a.b. All these notations convey the same meaning. Now consider a quantity 8abc. This is a result of multiplying 8, a, b and c. These quantities are referred to as the factors of the product 8abc. Also in this product, the number 8 is referred to as a coefficient of the remaining factors. In the broad sense of it even the quantity 'ab' is also referred to as a coefficient. But to differentiate it from a numerical coefficient, we refer to it as a literal coefficient. Also the product 8abc can be expressed as 8bac or 8cab. Only one prefers to present them in alphabetical order.

Now how do we express a quantity like a.a.a.a. We observe that 'a' has been multiplied by itself four times. The product of a quantity when multiplied by itself repeatedly is usually referred to as the power of that quantity and is expressed by writing the number of factors to the right of that quantity and above it. It will be like a^4. When a quantity is expressed in this form, the quantity 'a' is referred to as base and the numerical 4 is referred to as the power or index or exponent. As we treat a number without a sign as positive, any number whose power is not specified is understood to have its power as 1. Here one should understand the difference between the coefficient and power very clearly.

Usually the quantities like a^2 is read as 'a squared', and a^3 as 'a cubed' and so on. With this background we should be able to solve problems like finding the value of 5abc + 7d given that a = 2, b = 1, c = 3 and d = 3. It will be

5.2.1.3 + 7.3

That is, 30 + 21 = 51.

Examples 1

Find the values of the given expressions. Also given that a = 2, b = 3, c = 1, x = 2 and y = 3.

a. 8a + 5bc

= 8.2 + 5.3.1

= 16 + 15

= 31

b. $9a + c$

$= 9.2 + 1$

$= 18 + 1$

$= 19$

c. x^3

$= 2^3$

$= 2.2.2$

$= 8$

Now, what would be the value of 8abc, if one of the quantities a, b or c is zero. It will be zero and this is irrespective of other values. A factor which has its value as zero is called zero factor. Remember that every power of zero is zero.

Now is it that we only have quantities of the form 8ab or x^3. No, we often come across quantities like $\sqrt{ab}$ or $\sqrt[5]{bx^2}$. In mathematics, the $\sqrt{\ }$ sign is referred to as a radical sign and at this point let us define what square root is. The square root of any quantity is that value whose square is equal to the given expression. That is, $\sqrt{4} = 2$. If we square two we get four which is the required quantity under the radical sign. Similar to square root we have cube roots $\left(\sqrt[3]{\ }\right)$, the fourth $\left(\sqrt[4]{\ }\right)$ and the fifth $\left(\sqrt[5]{\ }\right)$, etc..... roots. Only that we have to multiply the given quantity the required number of times to get the quantity under the radical sign. Now we look at a couple of examples as to how to solve problems having the radical sign.

Examples 2

Find the value of

i. $\sqrt{\dfrac{9a^2}{100}}$

We know that square of 3a is $9a^2$ and square of 10 is 100. We write the given expression as $\left(\dfrac{9a^2}{100}\right)^{1/2}$.

Further $\left(\dfrac{9a^2}{100}\right)^{1/2} = \left(\dfrac{(3a)^2}{(10)^2}\right)^{1/2}$

$$= \frac{\left((3a)^2\right)^{1/2}}{\left((10)^2\right)^{1/2}} = \frac{(3a)^{2 \text{ x } 1/2}}{(10)^{2 \text{ x } 1/2}} = \frac{3a}{10}.$$

[**Note:** Here we have employed

i. $\left(\dfrac{x}{y}\right)^m = \dfrac{x^m}{y^m}$ and

ii. $(x^m)^n = x^{mn}$. [These laws will be studied later in the laws of indices.]

ii. $\sqrt[5]{\dfrac{32a^{10}}{243}}$

This can be expressed as $\left(\frac{32a^{10}}{243}\right)^{1/5}$.

Further

$$= \left(\frac{32a^{10}}{243}\right)^{1/5} = \frac{(32a^{10})^{1/5}}{(243)^{1/5}}$$

$$= \frac{(2^5)^{1/5} \text{ x } (a^{10})^{1/5}}{(3^5)^{1/5}} = \frac{2a^2}{3}$$

iii. $\sqrt[3]{\frac{a^2k^2}{3x^3}}$ if a = 8, k = 9, and x = 4.

We substitute the respective values

$$= \sqrt[3]{\frac{a^2k^2}{3x^3}} = \sqrt[3]{\frac{(8)^2.(9)^2}{3(4)^3}} = \sqrt[3]{27}$$

$$= (27)^{1/3} = (3^3)^{1/3} = 3^{3x1/3} = 3^1 = 3$$

Although we have been dealing with expressions till now, we have not come across expressions of the type: $8ab^2$, $-7ab^2$ and $9a^2bc$, $10a^2bc$. The terms which differ from each other only in terms of numerical coefficients but have the same literal coefficient part are referred to as like terms. Otherwise they are referred to as unlike terms. In accordance with this definition of like terms, the terms $8ab^2$, $-7ab^2$ and $9a^2bc$, $10a^2bc$ are like terms. In the first pair, we can observe that the quantity 'ab^2' is the same. Similarly in the second pair, the quantity a^2bc in both the expressions is the same. The next logical question would be: Are these expressions amenable to basic operations? Yes. But we start looking into these only after learning some fundamentals.

ADDITION AND SUBTRACTION OF LIKE TERMS AND UNLIKE TERMS

Addition of Like Terms with Same Signs

Case 1: Suppose we are given expressions like 3abc and 7abc and asked to compute their sum. If this is the case we should not worry much. Because adding like expressions with plus sign is as easier as adding positive numbers. Therefore, in this case add the numerals and suffix the common symbolic part which in this case happens to be abc. The sum will be, therefore,

3abc + 7abc = 10abc

Case 2: It is possible that all the like terms given may have minus (–) sign also. In this case also, we add the numerical coefficients and suffix the symbolic part to this numerical along with the minus sign. That is, if we are required to add –ab, –2ab and –2ab, we take the sum of numerical coefficients without their respective signs which will be 1 + 2 + 2 = 5. To this value, we suffix the common symbolic part (ab) along with the minus sign. That will be –5ab.

Subtraction of Like Terms with Same Signs

Suppose we are required to find the difference between 3abc and 7abc. We look at two scenarios. The value we would obtain by subtracting a larger quantity from the smaller quantity is not the same as one got by subtracting the lower value from the higher value. In the first scenario we have 7abc – 3abc which is 4abc. This is different from the value got in the second scenario which is 3abc – 7abc = – 4abc. Therefore, whenever a bigger value is subtracted from a smaller value, the process is as usual but to the solution part we prefix the minus sign, while the subtraction of a smaller value from a bigger value is nothing but the normal subtraction.

Addition of Unlike Terms

In this case, the first point we have to remember is that we do not get a single value when we add two or more terms which are unlike in nature. This certainly obviates the need for us to look at the result when we add or subtract unlike terms. That is, addition of –7abc and 3bc will be either –7abc + 3bc or 3bc – 7abc and the result if we wish to subtract these terms will be either –7abc – 3bc or 3bc + 7abc depending on whether we consider either 3bc or –7abc as the first or the second term.

MULTIPLICATION AND DIVISION OF LIKE TERMS AND UNLIKE TERMS

Multiplication of Two Like Terms with Same Signs

Case 1: Suppose we have two terms 7ab and 3ab. When we multiply these two terms, we get 7ab x 3ab = (7 x 3) a^{1+1} . b^{1+t} (Θ x^m . x^n = x^{m+n}) = $21a^2b^2$. The product of 7ab x 3ab will be the same as that of 3ab x 7ab. Irrespective of the order of multiplication, the product of two positive terms will be a positive term.

Similarly, the product of $8a^2b$ and $3a^2b$

= (8 x 3) a^2 . a^2 b . b

= 24 a^{2+2} b^{1+1}

= 24 a^4 b^2

Case 2: Suppose we have to compute the product of –7ab and –3ab, it will be equal to (–7 x –3) a^2b^2 = $21a^2b^2$, i.e. multiplication of two negative quantities gives us a positive quantity.

Multiplication of Two Like Terms with Opposite Signs

The product of –7ab and +3ab is (–7 x 3) a^2 b^2 = $–21a^2$ b^2. In other words, a term with minus sign when multiplied with a term having a positive sign, gives a product having a minus sign.

On the whole, one has to remember the following rules, while multiplying quantities,

+	x	+	=	+
–	x	–	=	+
–	x	+	=	–
+	x	–	=	–

(**Note:** The multiplication of binomials is shown below.)

Multiplication of Like Terms with Same Signs

Case 1: Suppose we are given 3ab and 8cd. The product of these two terms will be 3ab x 8cd = 24abcd. We would get the same result if 8cd is taken first instead of 3ab. Therefore, multiplication of two positive terms will give us a positive term.

Case 2: Suppose that we are asked to calculate the product of –3ab and –8cd. It will be –3ab x –8cd = 24abcd and an important point to remember is that the multiplication of two negative quantities gives us a positive quantity.

Multiplication of Two Unlike Terms with Opposite Signs

The product on multiplying – 4bc with 2a is – 8abc. That is, a term with minus sign multiplied with a term having a positive term gives a product which has a minus sign.

Division of Two Like Terms

Case 1: Suppose we have two terms 8ab and 4ab. On dividing the first by the second we have 8ab/4ab = 2 or 4ab/8ab = (1/2) depending on whether we consider either 8ab or 4ab as the first term. Irrespective of the order of division, the quotient of two positive terms will be a positive term.

Case 2: What will be the quotient if you divide –9ac by –3ac. We will get –9ac/–3ac = 3. In case of –3ac/–9ac, we will get 1/3. As in case 1, irrespective of the order of division, the quotient will be a positive term.

(**Note:** Observe that –9ac/–3ac is same as –9ac x 1/–3ac where 1/–3ac being the reciprocal of –3ac.)

We took monomials only for the sake of understanding the underlying principles in a better manner. However, these principles can be applied equally well to binomials, trinomials and polynomials also. The next part deals with them.

Now, do you find any difference between –7abc – 3bc and –7abc + (–3bc) and –7abc – (–3bc). The terms –7abc – 3bc and –7abc + (–3bc) are one and the same, the third expression is certainly different. The idea of introducing the bracket is to convey that the quantity within the brackets ought to be treated as a single quantity. Therefore, whenever one removes the brackets the changes necessary ought to be made specially with respect to the signs of the terms. Otherwise, you may end up with a wrong solution. And since we are mainly concerned with the multiplication aspect whenever brackets come into picture, we apply the same four rules we have seen while going through multiplication of like and unlike terms. The expression –7abc + (–3bc) on removal of the brackets will be –7abc + x –3bc. Since + x – gives us –, the expression will get simplified to –7abc – 3bc. The third expression –7abc – (–3bc) would be simplified to –7abc + 3bc on removal of the brackets. Regarding the change of signs whenever brackets are present, we make the following two important observations:

1. Whenever + sign precedes a bracket, the brackets can be removed without changing the signs of the elements within the brackets.
2. Whenever – sign precedes a bracket, they can be removed only by changing the signs of the elements within the brackets.

From these observations, we also conclude that if you want to introduce a + sign and include some of the terms of an expression in brackets, you can do so without changing the sign of the terms irrespective of them being + or –. That is, –7abc – 3ab can be written as –7abc + (–3ab) or 8a + bc can be written as 8a + (+bc). But if you want to introduce a – sign, more attention has to be paid. Every time you want to introduce a – sign, the signs of the terms to be included in the bracket has to be changed. In other words, a term which has a – sign should be changed to + sign and the term which has a + sign should be changed to – sign. Now let us look at a few examples as to how the basic operations are conducted in case of binomials, trinomials and polynomials.

Examples 3

i. Add 3a + 5b and 5a – 8b.

We position the binomials as shown below. This makes our computation part easier.

$$\begin{array}{rr} & 3a + 5b \\ (+) & 5a - 8b \\ \hline & 8a - 3b \\ \hline \end{array}$$

ii. Subtract 14bc – 5ac from 7bc + 9ac

$$\begin{array}{rr} & 7bc + 9ac \\ (-) & 14bc - 5ac \\ \hline & -7bc + 14ac \\ \hline \end{array}$$

Note that addition and subtraction are also performed on expressions like (1/2)a – (1/3)b and –a + (2/3)b (although we did not see expressions of this type above, they duly form a part of polynomials). In this case also we position the expressions as we did above. That will be

$$\begin{array}{ll} & \frac{1}{2}a - \frac{1}{3}b \\ (+) & -a + \frac{2}{3}b \\ \hline \end{array}$$

That is, we have to add (1/2) and –1. We follow the method first seen in addition of fractions. Therefore, the sum of 1/2 and –1 will be –1/2 and the sum of –1/3 and 2/3 will be 1/3. That is, the sum of the above two expressions will be (–1/2)a + (1/3)b.

The subtraction of the above two expressions will give us (3/2)a – b.

Multiplication of Binomials

To understand the multiplication of binomials, we should know what is meant by Distributive Law of Multiplication. Suppose that we are to multiply (a + b) and m. We treat (a + b) as a compound expression and m as a simple expression. Therefore, (a + b)m by definition will be:

= m + m + m + m + ... taken a + b times

= (m + m + m + ... taken a times)

+ (m + m + m + ... taken b times)

= am + bm

Similarly (a – b)m = am – bm and (a – b + c)m = am – bm + cm. This is referred to as Distributive Law of Multiplication and it says that the product of a compound expression by a simple expression is the algebraic sum of the partial products of each term of the compound expression by that simple expression.

In the above, if we write (c + d) in place of m, we will have

(a + b)(c + d) = a(c + d) + b(c + d)

= ac + ad + bc + bd

iii. Multiply (3a + d) and (b + c).

We employ (a + b)(c + d) = a(c + d) + b(c + d) = ac + ad + bc + bd. Therefore, (3a + d)(b + c) = 3a(b + c) + d(b + c) = 3ab + 3ac + bd + cd. (This procedure can be extended to trinomials and polynomials also.)

iv. Multiply 2a + 5c and 3d + 2b.

One way of doing this is to employ (a + b)(c + d) = ac + ad + bc + bd

That is,

(2a + 5c)(3d + 2b) = 2a(3d + 2b) + 5c(3d + 2b)

= 6ad + 4ab + 15cd + 10bc

In the second method, we position the binomials as we did in addition or subtraction and do the multiplication operation. That is,

$$\begin{array}{ll} & 2a + 5c \\ (x) & 3d + 2b \\ \hline & 6ad + 15cd \\ & \quad + 4ab + 10bc \\ \hline & 6ad + 15cd + 4ab + 10bc \\ \hline \end{array}$$

This product is the same as one obtained earlier.

v. Multiply $-\frac{1}{3}a - \frac{1}{4}b$ and $-\frac{2}{3}c + \frac{3}{4}e$

That is, we have to compute

$$(-\frac{1}{3}a - \frac{1}{4}b)\,(-\frac{2}{3}c + \frac{3}{4}e)$$

We write this as

$$-\frac{1}{3}a.(-\frac{2}{3}c + \frac{3}{4}e) - \frac{1}{4}b.(-\frac{2}{3}c + \frac{3}{4}e)$$

$$= (-\frac{1}{3}a)(-\frac{2}{3}c) + (-\frac{1}{3}a)(\frac{3}{4}e)$$

$$-(\frac{1}{4}b)(-\frac{2}{3}c) - (\frac{1}{4}b)(\frac{3}{4}e)$$

$$= \left(\frac{2}{9}\right)ac - \left(\frac{1}{4}\right)ae + \left(\frac{1}{6}\right)bc - \left(\frac{3}{16}\right)be.$$

(**Note:** While multiplying fractions, numerators and denominators of given fractions are multiplied respectively and the product also is expressed as a fraction.)

vi. Add 3ac + 5bd – 7cd and ac – 5bd – 4cd

$$\begin{array}{rr} & 3ac + 5bd - 7cd \\ (+) & ac - 5bd - 4cd \\ \hline & 4ac + 0 - 11cd \end{array}$$

vii. Multiply 3a + 5b – 7d and c – 4e – 5

That is, we require (3a + 5b – 7d) x (c – 4e – 5)

= 3a (c – 4e – 5) + 5b (c – 4e – 5) – 7d(c – 4e – 5)

= 3ac – 12ae – 15a + 5bc – 20be – 25b – 7cd + 28de + 35d.

DIMENSIONS AND DEGREE OF AN EXPRESSION

Binomials, Trinomials and Polynomials which we have seen above are not the only type. We can have them in a single variable say, 'x' and of the form $x^2 + 4x^2 – 5x + 1$. Before we go into these, first let us understand what is meant by dimension. It is defined as each of the letters (symbols) comprising a term. That is, in term abc, a, b and c are the three letters which indicate that the dimension is 3. Compared to this, we define degree as number of letters in a term. The number of letters in the term abc are 3 and therefore, it can be said that it is of 3rd degree. You find these two somewhat similar. However, the degree of an expression consisting two or more terms is of that term which has the highest dimension, that is, in an expression $6a^3x^3 + 5b^2x^5 – 3c^2x^2$, we find that the dimensions of the term $6a^3x^3$ is 6, that of the term $5b^2x^5$ is 7 and that of the term $3c^2x^2$ is 4. The degree of this expression is, therefore, 7. An expression in which the terms have same dimensions is said to be a homogeneous expression. The like and unlike terms which we have seen earlier is, at the most, of two degrees. We do have terms of higher degrees also. An expression of the form $3x^5 + 7x^4 + x^3 – 2x + 5$ is of degree 5. In polynomials we deal with expressions like these. In this part, first we look at their addition, subtraction and multiplication. In case of division, we will list the steps and look at a couple of examples involving binomials before we go to polynomials. In the latter part we look at how to factorize expressions of any given degree.

ADDITION

Example 4

Add $4x^4 + 3x^3 – x^2 + x + 6$ and $–7x^4 – 3x^3 + 8x^2 + 8x – 4$

We write them one below the other as shown below.

$$\begin{array}{ll} & 4x^4 + 3x^3 - x^2 + x + 6 \\ (+) & -7x^4 - 3x^3 + 8x^2 + 8x - 4 \\ \hline & -3x^4 + 0 + 7x^2 + 9x + 2 \end{array}$$

Example 5

Add $5x^5 – 6x^3 + 4x^2 + 3x – 7$, $3x^5 – 2x^4 + 3x^2 + 6x – 1$ and $– 3x^4 + x^3 – 5x^2 + 7x + 4$

$$\begin{array}{l} 5x^5 + -6x^3 + 4x^2 + 3x - 7 \\ 3x^5 - 2x^4 + + 3x^2 + 6x - 1 \\ -3x^4 + x^3 - 5x^2 + 7x + 4 \\ \hline 8x^5 - 5x^4 - 5x^3 + 2x^2 + 16x - 4 \end{array}$$

Example 6

Subtract $–7x^4 – 3x^3+ 8x^2 + 8x – 4$ from $4x^4 + 3x^3 – x^2 + x + 6$

$$\begin{array}{ll} & 4x^4 + 3x^3 - x^2 + x + 6 \\ (-) & -7x^4 - 3x^3 + 8x^2 + 8x - 4 \\ \hline & 11x^4 + 6x^3 - 9x^2 - 7x + 10 \end{array}$$

In this problem, in some expressions we do not find terms of certain powers. They have been left as blanks.

Example 7

Subtract the first from the second and sum the difference with the third expression. The expressions are given below.

$5x^5 – 6x^3 + 4x^2 + 3x – 7$, $3x^5 – 2x^4 + 3x^2 + 6x – 1$ and $–3x^4 + x^3 – 5x^2 + 7x + 4$.

The difference of first two expressions is given by

$$\begin{array}{ll} & 3x^5 - 2x^4 + 3x^2 + 6x - 1 \\ (-) & 5x^5 - 6x^3 + 4x^2 + 3x - 7 \\ \hline & -2x^5 - 2x^4 + 6x^3 - x^2 + 3x + 6 \end{array}$$

The sum of the difference between the first two expressions and the third expression will be as shown below.

$$\begin{array}{ll} & -2x^5 - 2x^4 + 6x^3 - x^2 + 3x + 6 \\ (+) & -3x^4 + x^3 - 5x^2 + 7x + 4 \\ \hline & -2x^5 - 5x^4 + 7x^3 - 6x^2 + 10x + 10 \end{array}$$

MULTIPLICATION

Example 8

Multiply $3x^5 + 4x^3 + 2x – 1$ and $x^4 + 2x^2 + 4$.

The product is given by

$3x^5 . (x^4 + 2x^2 + 4) + 4x^3. (x^4 + 2x^2 + 4) + 2x .$

$(x^4 + 2x^2 + 4) – 1 . (x^4 + 2x^2 + 4)$

$= 3x^5 . x^4 + 3x^5 . 2x^2 + 3x^5 . 4 + 4x^3 . x^4 + 4x^3 .$

$2x^2 + 4x^3 . 4 + 2x . x^4 + 2x . 2x^2 + 2x . 4 – x^4 – 2x^2 – 4$

To simplify the above we employ a rule which we will learn in laws of indices. It states that $x^m \cdot x^n = x^{m+n}$

$= 3x^9 + 6x^7 + 12x^5 + 4x^7 + 8x^5 + 16x^3 + 2x^5 + 4x^3 + 8x - x^4 - 2x^2 - 4$

Now we collect like terms and simplify them. We obtain $3x^9 + 10x^7 + 22x^5 - x^4 + 20x^3 - 2x^2 + 8x - 4$.

DIVISION

Before taking up division of polynomials, let us acquaint ourselves with some basics. Suppose we are asked to divide 16 by 2. We know that on dividing 16 by 2 we get 8. In mathematics we call 16, 2 and 8 by specific names – 16 is called dividend, 2 is called the divisor and 8 is the quotient. However, it is not always that we get an integer like 8 when we divide a number by another. For instance, divide 9 by 2. In addition to the dividend (9), divisor (2) and quotient (4) we are left with another term 1. This is referred to as the remainder. When the dividend is not exactly divisible by the divisor we get a remainder. We find these terms even when one expression is divided by another. Also we follow these rules.

1. We arrange the terms of the divisor and the dividend in ascending or descending powers of some common letter. Ascending order refers to arranging terms from lower power to higher powers and descending orders refers to the opposite of this. Usually we write them in the descending order.
2. Divide the term on the left of the dividend by the term left of the divisor and put the result in the quotient.
3. Multiply the whole divisor by this number (quotient) and put the resultant product under the dividend.
4. Subtract the product from the dividend and bring down the required number of terms as may be deemed necessary.
5. Repeat this procedure until all the terms in the dividend have been brought down.

We understand this with the help of a couple of examples.

Example 9

Divide $x^2 + 4x + 4$ by $x + 2$.

We find that the terms of the dividend ($x^2 + 4x + 4$) and the divisor ($x + 2$) are already in the descending order. The left most term in the dividend is x^2, while in the divisor it is x. We find the quotient as

$$\frac{x^2}{x} = \frac{x \cdot x}{x} = x.$$

We multiply the divisor $x + 2$ with this quotient x. We get $x^2 + 2x$. We write this under the dividend as shown below.

$$\begin{array}{lll} x + 2\ ) & x^2 + 4x + 4 & (\ x + 2 \\ (-) & x^2 + 2x & \\ \hline & 2x + 4 & \\ (-) & 2x + 4 & \\ \hline & 0 & \end{array}$$

On subtracting $x^2 + 2x$ from the dividend we obtain $2x + 4$. ($x^2 + 4x + 4 - (x^2 + 2x)$

$= x^2 + 4x + 4 - x^2 - 2x$)

We write this expression as shown above.

At this stage, we take the left most quantity of the difference (dividend – product) and that of the divisor and obtain their quotient. It will be

$$\frac{2x}{x} = 2$$

Since the sign of the quotient is positive we write it as shown. Then we multiply x + 2 with 2. That will be 2x + 4. We write under the difference got earlier and subtract it from the difference. We get 2x + 4 – (2x + 4) = 2x + 4 – 2x – 4 = 0. This is shown in the example above. Since the dividend is exactly divisible by the divisor the remainder is zero.

After solving this problem can we say that x + 2 is a factor of $x^2 + 4x + 4$? Of course we can. As we write 8 = 2.4 or 1.8, we can write

$$x^2 + 4x + 4 = (x + 2)(x + 2)$$

(**Note:** Division of expressions where some of the terms are fractions is also carried out in the same manner we have seen above.)

Example 10

Divide $6x^5 – x^4 + 4x^3 – 5x^2 – x – 15$ by $2x^2 – x + 3$. The working of this example and the various steps constituting it are stated below.

$2x^2 – x + 3)\ 6x^5 – x^4 + 4x^3 – 5x^2 – x – 15\ (3x^3 + x^2 – 2x – 5$

(–)	$6x^5 – 3x^4 + 9x^3$
	$+ 2x^4 – 5x^3 – 5x^2$
(–)	$2x^4 – x^3 + 3x^2$
	$– 4x^3 – 8x^2 – x$
(–)	$– 4x^3 + 2x^2 – 6x$
	$–10x^2 + 5x – 15$
(–)	$–10x^2 + 5x – 15$
	0

Step 1: $\frac{6x^5}{2x^2} = 3x^3$. Take this ratio to the quotient part.

Step 2: Multiply $2x^2 – x + 3$ by $3x^3$. We obtain $6x^5 – 3x^4 + 9x^3$. We write this under the dividend as shown and then subtract it from the dividend. We get $2x^4 – 5x^3 – 5x^2$ which is written as shown in the example. Observe that the difference part is only $2x^4 – 5x^3$. The quantity $–5x^2$ is brought down from the dividend (remember the point 4 in steps).

Step 3: We again take $2x^4/2x^2 = x^2$. We multiply the divisor with this quantity. We obtain $2x^4 – x^3 + 3x^2$ and write this under the difference obtained earlier. Subtract them. We get $– 4x^3 – 8x^2$. As earlier, the quantity –x is brought down from the dividend. Why we have to bring necessary quantities down from the dividend should be clear by now.

Step 4: We repeat the procedure again. This time it will be $\frac{-4x^3}{2x^2} = –2x$.

On multiplying this with the divisor we obtain $– 4x^3 + 2x^2 – 6x$. This, when written under the difference and subtracted, we obtain $– 10x^2 + 5x – 15$.

Step 5: The ratio we obtain is $\frac{-10x^2}{2x^2} = –5$. This when multiplied with the divisor yields $–10x^2 + 5x – 15$. This written under the earlier difference and subtracted gives us 0. Since the dividend is exactly divisible by the divisor we obtain the remainder as 0. Thus $(2x^2 – x + 3)$ and $(3x^3 + x^2 – 2x – 5)$ are the factors of $6x^5 – x^4 + 4x^3 – 5x^2 – x – 15$.

Factorization of Expressions

Above we have seen that $(2x^2 – x + 3)$ and $(3x^3 + x^2 – 2x – 5)$ are the factors of $6x^5 – x^4 + 4x^3 – 5x^2 – x – 15$. In this case, we are able to find one factor given the other one. How are we going to solve in case when we are not given either of them. Finding the factors of a given expression forms the part of our attention now. First, we look at binomial expressions and once we understand this we move on to trinomials and polynomials. If the given expression is in the form of an identity (we look at them shortly) our job becomes easier; otherwise, we have to adopt trial and error method until we get at least one of the factors. Once we know one of the factors then by employing the division method we can get other factors.

Example 11

Factorize $x^2 + 6x + 9$.

If we substitute x = 1, the value of the expression will be $(1)^2 + 6(1) + 9 = 16$

Since the value of the numerical expression is not 0, we substitute another value. We will continue to do so until we get a zero.

If we substitute x = –1, the value of the expression will be $(–1)^2 + 6(–1) + 9 = 4$

If we substitute x = 2, the value of the expression will be $(2)^2 + 6(2) + 9 = 25$

If we substitute x = –2, the value of the expression will be $(–2)^2 + 6(–2) + 9 = 1$

If we substitute x = 3, the value of the expression will be $(3)^2 + 6(3) + 9 = 36$

We substitute x = –3, the value of the expression will be $(–3)^2 + 6(–3) + 9 = 0$

For x = –3, the value of the expression is 0. That is, x + 3 is one of the factors of the expression $x^2 + 6x + 9$. To obtain the other factor we divide the expression by the factor we obtained. That will be

$$
\begin{array}{lll}
x + 3\ ) & x^2 + 6x + 9 & (\ x + 3 \\
(–) & x^2 + 3x & \\
\hline
 & 3x + 9 & \\
(–) & 3x + 9 & \\
\hline
 & 0 &
\end{array}
$$

From the division, we observe that x + 3 is the other factor. When this is equated to zero we obtain x = – 3. Therefore, the factors of $x^2 + 6x + 9$ are $(x + 3)(x + 3)$ or $(x + 3)^2$.

In the above example we note that $x^2 + 6x + 9 = (x + 3)^2$. Isn't this identical to $a^2 + 2ab + b^2 = (a + b)^2$? The value of a being x and that of b equal to 3. This is one of the basic identities we get to see at in algebra. We come across the others as we look at other examples.

Example 12

Factorize $x^2 – 4x + 4$.

If we substitute x = 1, the value of the expression will be $(1)^2 – 4(1) + 4 = 1$

If we substitute x = –1, the value of the expression will be $(–1)^2 – 4(–1) + 4 = 9$

If we substitute x = 2, the value of the expression will be $(2)^2 – 4(2) + 4 = 0$

For x = 2, the value of the expression is 0. That is, x – 2 (observe that x – 2 = 0 and x = 2 are one and the same) is one of the factors of the expression $x^2 – 4x + 4$. To obtain the other factor we divide the expression by the factor we got.

That will be

$$
\begin{array}{lll}
x-2\ ) & x^2-4x+4 & (\ x-2 \\
\quad (-) & x^2-2x & \\
\hline
 & -2x+4 & \\
\quad (-) & -2x+4 & \\
\hline
 & 0 &
\end{array}
$$

From the division we observe that $x - 2$ is the other factor. When this is equated to zero we obtain $x = 2$. Therefore, the factors of $x^2 - 4x + 4$ are $(x - 2)(x - 2)$ or $(x - 2)^2$.

Now, we look at another identity which is similar to the one you have seen earlier except the (–) sign. The identity is $(a - b)^2 = a^2 - 2ab + b^2$. The advantage of being familiar with identities is that you do not have to sweat it out by factorizing each and every expression you are given. On the other hand, it is not mandatory that each and every expression given should be in conformation with some identity. In this case there is no easy way out except solving the problem by trial and error method to start with and then go for division in order to know other factors.

Another identity of second degree we often come across is

$$a^2 - b^2 = (a + b)(a - b)$$

According to this identity the difference of squares of any two quantities is equal to the product of the sum and the difference of the two quantities.

We take an example and look at how to obtain the factors when expressions similar to the above identity are given.

Example 13

Factorize $4p^2 - 25$.

We can also write this expression as $(2p)^2 - (5)^2$. In turn $(2p)^2 - (5)^2$ can be written as $(2p + 5)(2p - 5)$. Therefore, the factors of the given expression are $(2p + 5)$ and $(2p - 5)$.

Example 14

Factorize $a^4 - 81$.

We write the given expression as $(a^2)^2 - (9)^2$. This can be further expressed as $(a^2 + 9)(a^2 - 9)$. Can we simplify this further? Yes, observe that after applying the priniciple once, we apply the same for $a^2 - 9$. Now, $a^2 - 9$ can be expressed as $(a - 3)(a + 3)$. Therefore, the factors of the given expression are $(a^2 + 9)$, $(a + 3)$ and $(a - 3)$.

Let us look at some of the other identities in addition to what we have seen above and take up some examples.

$$
\begin{aligned}
(a + b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3 \\
(a - b)^3 &= a^3 + 3a^2(-b) + 3a(-b)^2 + (-b)^3 \\
&= a^3 - 3a^2b + 3ab^2 - b^3 \\
a^3 + b^3 &= (a + b)(a^2 - ab + b^2) \\
a^3 - b^3 &= (a - b)(a^2 + ab + b^2).
\end{aligned}
$$

Example 15

Factorize $27x^3 + 54x^2 + 36x + 8$.

The given expression can be written as $(3x)^3 + 3(3x)^2(2) + 3(3x)(2)^2 + (2)^3$.

We find this similar to $a^3 + 3a^2b + 3ab^2 + b^3$ which is $(a + b)^3$. In place of a and b, we have 3x and 2 respectively. Therefore, the factors of the given expression are $(3x + 2)(3x + 2)(3x + 2)$ or $(3x + 2)^3$.

Example 16

Factorize $64x^3 - 48x^2 + 12x - 1$.

The given expression can be written as $(4x)^3 + 3(4x)^2(-1) + 3(4x)(-1)^2 - (1)^3$.

This, when simplified, gives us $64x^3 - 48x^2 + 12x - 1$. We find this similar to $a^3 - 3a^2b + 3ab^2 - b^3$, which is $(a - b)^3$. In place of a and b, we have 4x and 1 respectively. Therefore, the factors of the given expression are $(4x - 1)(4x - 1)(4x - 1)$ or $(4x - 1)^3$.

Example 17

Factorize $x^3 + 3x^2 + 3x + 1$.

If we substitute x = 1, the value of the expression will be $(1)^3 + 3(1)^2 + 3(1) + 1 = 8$

If we substitute x = –1, the value of the expression will be $(-1)^3 + 3(-1)^2 + 3(-1) + 1 = 0$

We observe that when x = –1, the value of the expression will be 0. Therefore, x = –1 or x + 1 is one of the factors of the given expression. To simplify this further, we divide the expression $x^3 + 3x^2 + 3x + 1$ by x + 1. That will be

x + 1)	$x^3 + 3x^2 + 3x + 1$	($x^2 + 2x + 1$
(–)	$x^3 + x^2$	
	$2x^2 + 3x$	
(–)	$2x^2 + 2x$	
	$x + 1$	
(–)	$x + 1$	
	0	

We observe that $x^2 + 2x + 1$ is the other factor. Since $x^2 + 2x + 1$ is of second degree we have to check whether it is possible to factorize it further or not. We observe that the expression is identical to $(a + b)^2 = a^2 + 2ab + b^2$. In place of a we have x and in the place of b we have 1. Therefore, $x^2 + 2x + 1 = (x + 1)^2$, that is, the expression $x^3 + 3x^2 + 3x + 1$ can be factorized into $(x + 1)(x + 1)(x + 1)$ or $(x + 1)^3$. This gives us the identity $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$ which we have seen above.

That is, similar to $(a + b)^2$ and $(a - b)^2$, we have $(a + b)^3$ and $(a - b)^3$. In fact, we have identities like these for 4th, 5th, 6th and so on .... powers. As you can observe, the expansion of $(a + b)^3$ has four terms in it and an identity with fourth power has five terms. Therefore, the number of terms will go on increasing as we increase the power of an identity and computing this may prove to be a tedious job.

Example 18

Factorize $64x^3 - 27y^3$.

The given expression can be written as $(4x)^3 - (3y)^3$. From above we know that the factors of $(a)^3 - (b)^3 = (a - b)(a^2 + ab + b^2)$. Therefore, the factors of $(4x)^3 - (3y)^3 = (4x - 3y)((4x)^2 + (4x)(3y) + (3y)^2)$. On simplifying, we obtain $(4x - 3y)(16x^2 + 12xy + 9y^2)$.

Example 19

Factorize $a^3 b^3 + 512$.

The given expression can be written as $(ab)^3 + (8)^3$. From above we know that the factors of $(a)^3 + (b)^3 = (a + b)(a^2 - ab + b^2)$. Therefore, the factors of $(ab)^3 + (8)^3 = (ab + 8)((ab)^2 - (ab)(8) + (8)^2)$. On simplifying we obtain $(ab + 8)(a^2b^2 - 8ab + 64)$.

Now let us look at example where the expression is of the fourth degree.

Example 20

Factorize $9x^4 - 12x^3 - 2x^2 + 4x + 1$.

If we substitute x = 1, the value of the expression will be

$$9(1)^4 - 12(1)^3 - 2(1)^2 + 4(1) + 1 = 0$$

That is, x = 1 or x – 1 = 0 is one of the factors of this expression. The other factors are obtained by dividing the expression with x – 1.

$$\begin{array}{rl} x - 1) & 9x^4 - 12x^3 - 2x^2 + 4x + 1 \quad (9x^3 - 3x^2 - 5x - 1 \\ (-) & 9x^4 - 9x^3 \\ \hline & -3x^3 + 2x^2 \\ (-) & -3x^3 + 3x^2 \\ \hline (-) & -5x^2 + 4x \\ & -5x^2 + 5x \\ \hline & -x + 1 \\ (-) & -x + 1 \\ \hline & 0 \\ \hline \end{array}$$

The other factor is $9x^3 - 3x^2 - 5x - 1$. To get other factors, we substitute x = 1 in the above factor. The value of the expression will be

$$9(1)^3 - 3(1)^2 - 5(1) - 1 = 0$$

That is, x – 1 = 0 or x = 1 is again a factor of $9x^3 - 3x^2 - 5x - 1$. We again divide.

$$\begin{array}{rl} x - 1) & 9x^3 - 3x^2 - 5x - 1 \quad (9x^2 + 6x + 1 \\ (-) & 9x^3 - 9x^2 \\ \hline & 6x^2 - 5x \\ (-) & 6x^2 - 6x \\ \hline & x - 1 \\ (-) & x - 1 \\ \hline & 0 \\ \hline \end{array}$$

That is, $9x^2 + 6x + 1$ is another factor. We find that $9x^2 + 6x + 1$ is identical to $a^2 + 2ab + b^2$. That is, for a = 3x and b = 1, we have $9x^2 + 6x + 1 = (3x + 1)^2$. Therefore, the factors of $9x^4 - 12x^3 - 2x^2 + 4x + 1$ are $(x - 1)(x - 1)(3x + 1)^2$ or $(x - 1)^2 (3x + 1)^2$.

Second Method to Solve Binomials of Second Degree

In this part, we look at another method to obtain the factors of an expression. In the above you have seen that $x^2 - 4x + 4 = (x - 2)^2$ or $(x - 2)(x - 2)$. If you observe it carefully we find that the middle number – 4 is the sum of –2 and –2 and the last term 4 is the product of –2 and –2. We take another example. You are given $x^2 + 15x + 56$ and asked to factorize it. Now if you think that, say, 6 and 7 are the factors of this expression then their product should be equal to 56 and their sum should be equal to 15. However in this case, we observe that the product is 42 and the sum is 13. Therefore, 6 and 7 cannot be the factors of this expression. Now try 7 and 8. We find that their product is 56 and the sum 15. That is, 7 and 8 are the factors of the given expression. This can be clarified by multiplying (x + 8) and (x + 7). One point to which we have to pay attention is that we have to take even

signs into consideration. For instance, consider an expression $x^2 - 17x + 70$. What could be the factors of this expression? Let us try 7 and 10. No doubt, the product is 70 and the sum 17. Still these cannot be the factors of the given expression, because the sum is –17 and we got only 17. Now let us try –7 and –10. The sum of these two numbers gives us –17 and their product as 70. This is what we require. Therefore, the factors are $x - 7$ and $x - 10$ (observe that in this case if we took $x = -7$ and $x = -10$, we would have got the factors as $x + 7$ and $x + 10$, whose multiplication would give us $x^2 + 17x + 70$ and not $x^2 - 17x + 70$. That is, the values should be considered as they are). Now let us consider an expression $x^2 - 3x - 70$. Let us try 7 and –10 for this expression. The sum of these two values is $-10 + 7 = -3$ and the product being –70. That is, $x + 7$ and $x - 10$ are two factors of the given expression and not $x - 7$ and $x + 10$.

SUMMARY

- The collection of alphabets in algebra are referred to as terms and the collection of these terms is called an algebric expression. The algebric expression can be either simple or compound.
- The compound expression containing two terms is called “abinomial” and that containing three terms is called a “trinomial” and that containing more than three terms is called “polynomial”. These can be represented with similar signs or opposite signs. They can be added, subtracted, multiplied or divided.

Lesson 3

Simultaneous and Quadratic Equations

After reading this lesson, you will be conversant with:

- Solving Simultaneous Equations
- Solving Quadratic Equations

SOLVING SIMULTANEOUS EQUATIONS

Before we look at simultaneous equations, let us brush up some of the fundamentals. First, we will define what is meant by an equation; it is a statement which indicates that two algebraic expressions are equal. For instance, let $3x - 4$ be an expression and $5x - 10$ be another expression. If these two expressions are related to each other by an equality sign in the fashion shown below we call it as an equation.

$$3x - 4 = 5x - 10 \quad \text{..... (1)}$$

The side on which we have the expression $3x - 4$ is referred to as Left Hand Side (LHS) and the one which has $5x - 10$ as the Right Hand Side (RHS). If we substitute $x = 3$ in the above equation we find that both sides of the equation give us 5. Now we substitute some other value say $x = 2$. We find that the LHS gives us 2 whereas the RHS gives us 0. Looking at these two cases we conclude that only when $x = 3$, the equation holds and for other values of x it does not. But consider an equation which is shown below.

$$3x + 2 + 2x - 5 = 5x - 3 \quad \text{..... (2)}$$

The LHS and the RHS of this equation give us the same values for any value of x. In other words, this equation holds for any value of x. Equations like these are called identities and the one we have seen before the above are referred to as equations of condition or more simply as a equations.

Above we have seen that only when we have substituted $x = 3$ in equ.(1), it holds true. That is, the value of $x = 3$ is said to be satisfying the equation. Since we are expected to find the value of x for which the equation holds true, the quantity x is known as the unknown quantity. The value of x found after solving the equation is called the solution or the root of the equation.

While solving equations, we have to remember these points.

1. If we are to add or subtract any quantity from/to one side of the equation, we should do so for the other side also. We look at this by taking an example.

   For instance, we are required to solve the equation

   $$x + 3 = 15$$

   That is, on the LHS we ought to have only x. How can we go about this job? Can we subtract 3 from the LHS so that +3 and –3 cancel each other leaving behind only x? We can. But as stated above, this operation should be done on both sides of the equation. That is, we will have

   $$x + 3 - 3 = 15 - 3$$
   $$x + 0 = 12$$
   $$x = 12$$

   If we do not perform this operation on both the sides, the balance which exists between the sides gets disturbed; as a result, the equality sign loses its relevance and thereby has no meaning.

   We take another example and check the same for addition. We have an equation $x - 3 = 12$, for which we have to obtain a solution.

   $$x - 3 + 3 = 12 + 3$$
   $$x = 15$$

   As we are aware of this, while solving equations we directly **transpose** the quantity to the other side of the equation with its sign changed. Here we introduced a new word “Transpose”. What is meant by Transposing? Bringing any term from one side of the equation to the other side is called transposing.

2. If we are to multiply or divide a particular element or the whole expression on one side of the equation, then we should do the same on the other side of the equation also. Let us take an example and understand this. We have to find the solution for the equation 3x + 5 = 20. We begin by subtracting 5 from both the sides. That will be

$$\begin{aligned} 3x + 5 - 5 &= 20 - 5 \\ 3x + 0 &= 15 \\ 3x &= 15 \end{aligned}$$

Since only x ought to be there on the LHS (i.e. solving for x), we divide the LHS by 3 and do a similar operation on the RHS also. We have

$$\frac{3x}{3} = \frac{15}{3}$$

$$\begin{aligned} 1.x &= 5 \\ x &= 5 \end{aligned}$$

Therefore, x = 5 is the solution of the given equation.

Suppose we are given an equation like

$$\frac{x-4}{3} = 6$$

and asked to solve, how should we proceed?

We begin by multiplying both the sides of the equation by 3. We have

$$\begin{aligned} \frac{x-4}{3} \times 3 &= 6 \times 3 \\ x - 4 &= 18 \\ x - 4 + 4 &= 18 + 4 \\ x &= 22. \end{aligned}$$

Example 1

On some occasions, we are asked to solve equations like

$$\frac{3x-2}{5} + \frac{1}{3} = \frac{x}{3} + \frac{x}{4}$$

In problems like these, we collect the x terms (bringing x terms to one side and the constants to the other side). Then, we have

$$\frac{3x-2}{5} - \frac{x}{3} - \frac{x}{4} = \frac{-1}{3}$$

We take the LCM (Least Common Multiple) of 5, 3 and 4. It is 60.

$$\frac{12(3x-2) - 20x - 15x}{60} = \frac{-1}{3}$$

$$\frac{x-24}{60} = \frac{-1}{3}$$

This can also be expressed as

$$x - 24 = 60 \times \frac{-1}{3} = -20$$

Therefore, $x = 24 - 20 = 4$

That is, x = 4 is the root of the given equation.

Till now we have considered a single equation and looked at how to obtain its solution. In this part, we look at simultaneous equations which generally contain two or more than two variables and also learn how to solve them by four different methods. You will find that the fourth 'Diagonalization method' is quite different from the first three.

Method 1

Above we have seen equations wherein we are required to find the value of the variable x only. Apart from the equations of the type we have seen above, simple equations of the form $7x + 2y = 47$ also exists. As against the equations $x – 3 = 12$ or $3x + 5 = 20$ where we had a single variable x, we have two variables x and y. If we express this equation in terms of y it will be

$$\begin{aligned} 7x + 2y &= 47 \\ 7x – 7x + 2y &= 47 – 7x \\ 0 + 2y &= 47 – 7x \\ 2y &= 47 – 7x \\ \frac{2y}{2} &= \frac{47 – 7x}{2} \\ y &= \frac{47 – 7x}{2} \end{aligned}$$

In this expression, by giving values to x and solving the expression we get corresponding values for y.

Now if another equation of the form $5x – 4y = 1$ is expressed in terms of y, we have

$$y = \frac{5x – 1}{4}$$

For this equation also if we substitute values for x, we would get corresponding values for y.

At this stage if we want values which satisfy both the equations then the values of y in both these equations should be identical. That is,

$$\frac{47 – 7x}{2} = \frac{5x – 1}{4}$$

On cross multiplying, we have $4(47 – 7x) = 2(5x – 1)$

$$188 – 28x = 10x – 2$$

Collecting all x terms and constants, we have

$$\begin{aligned} –10x – 28x &= – 2 – 188 \\ –38x &= – 190 \\ x &= 5 \end{aligned}$$

We substitute this value of $x = 5$, in either of equations to get the value of y.

$$\begin{aligned} 5x – 4y &= 1 \\ 5(5) – 4y &= 1 \\ – 4y &= 1 – 25 \\ – 4y &= – 24 \\ y &= –24/ – 4 = 6. \end{aligned}$$

Therefore, the common values of x and y which satisfy these two equations simultaneously are $x = 5$ and $y = 6$. Therefore, we can define simultaneous equations as two or more equations which are satisfied by the same values of x and y (unknown quantities). Is this the only method to solve simultaneous equations? No, we have more methods which we employ in solving simultaneous equations.

Method 2

In this method we eliminate either x or y, get the value of other variable and then substitute that value in either of the original equations to get the value of the other variable. Let us look at it.

We are given two equations $3x + 4y = 10$ and $4x + y = 9$. We write them as follows.

$$3x + 4y = 10 \quad \ldots (1)$$

$$4x + y = 9 \quad \ldots (2)$$

In this example, let us eliminate y. Multiply equ. (2) by 4. We have

$$4(4x + y = 9)$$

which is $16x + 4y = 36 \quad \ldots (3)$

We observe that the coefficients of y in equations (1) and (3) are one and the same, and therefore, we subtract equ.(1) from equ.(3).

$$\begin{array}{r} 16x + 4y = 36 \\ -(3x + 4y = 10) \\ \hline 13x + 0 = 26 \\ \hline \end{array}$$

At this point one should remember that the signs of equation (1) should be changed before subtracting.

$$13x = 26$$

$$x = 26/13 = 2$$

We now substitute the value of $x = 2$ in either equ. (1) or (2). Let us substitute in equ. (2).

$$4x + y = 9$$

$$4(2) + y = 9$$

$$8 + y = 9$$

$$y = 9 - 8 = 1$$

That is, the values of x and y for which both the equations are satisfied are $x = 2$ and $y = 1$.

We substitute these values in equ. (1) and check.

$$3x + 4y = 10$$

$$3(2) + 4(1) = 10$$

$$6 + 4 = 10$$

$$10 = 10$$

That is, LHS = RHS.

Method 3

In this method we express one equation in terms of either x or y and then substitute it in the other equation. On simplifying this equation and solving it, we get the value of one variable which is then substituted in one of the original equations to get the value of the other variable. Let us take an example.

Example 2

Solve $x + 2y = 13$ and $3x + y = 14$

We have $x + 2y = 13 \quad \ldots (1)$

and $3x + y = 14 \quad \ldots (2)$

We express equ.(1) in terms of x. It will be x = 13 – 2y. We substitute this in equ.(2).

We have, $3(13 - 2y) + y = 14$

$39 - 6y + y = 14$

$39 - 5y = 14$

$-5y = 14 - 39$

$-5y = -25$

$y = -25/-5$

$y = 5$

We substitute this value of y = 5 in equ. (1). It will be

$x + 2(5) = 13$

$x + 10 = 13$

$x = 13 - 10 = 3$

Therefore, the values of x = 3 and y = 5 satisfy both the equations.

Method 4

This method is also employed in Simplex method which we will come across while learning Linear Programming.

This method creates a series of 1s across the diagonal, and 0s elsewhere. It is explained through the examples below.

Example 3

Solve

$2x - y - z = 2$

$x + y + z = 1$

$x + y + 2z = 1$

We may represent the above system of equations by the table below. For example, the last row of the table reads: 1x + 1y + 2z = 1, which is exactly the last equation given above.

The first row and the first column are merely for reference. We will therefore number the rows and columns from 0 onwards. Therefore, the third equation is represented by the third row, columns 1 to 4.

Table 0

EQUATION	x	y	z	RHS
1	2	–1	–1	2
2	1	1	1	1
3	1	1	2	1

The idea is to progressively introduce 1 in the diagonal cells (row 1 and column 1, row 2 and column 2, row 3 and column 3) and 0s elsewhere using only the following operations.

- Multiply a row by any constant.
- Add a multiple of any row to any other row.

We introduce 1 in cell (1, 1) and 0s in the remaining cells of column 1 by multiplying row 1 by 0.5 and then subtracting the modified row 1 from rows 2 and 3.

That is,

Row 1: Row 1 x 0.5

Row 2: Row 2 + (–1 x modified Row 1)

Row 3: Row 3 + (–1 x modified Row 1).

Table 1

EQUATION	x	y	z	RHS
1	1.0	– 0.5	– 0.5	1.0
2	0.0	1.5	1.5	0.0
3	0.0	1.5	2.5	0.0

Now we want to introduce 1 in cell (2, 2) and 0s in the other cells of column 2. To achieve this, we need to multiply row 2 by (1/1.5). Follow this by adding 0.5 times row 2 to row 1, and adding –1.5 times row 2 to row 3.

Table 2

EQUATION	x	y	z	RHS
1	1.0	0.0	0.0	1.0
2	0.0	1.0	1.0	0.0
3	0.0	0.0	1.0	0.0

We now introduce 1 in cell (3, 3) and 0s elsewhere in column 3. But note that cell (3, 3) is 1 and cell (1, 3) is 0 already. To achieve our objective, we need cell (2, 3) = 0. This is done by adding – 1 times row 3 to row 2.

Table 3

EQUATION	x	y	Z	RHS
1	1.0	0.0	0.0	1.0
2	0.0	1.0	0.0	0.0
3	0.0	0.0	1.0	0.0

Since we cannot go any further in the Diagonalization process, let us construct the equations implied by Table 3.

$1x + 0y + 0z = 1.0$

$0x + 1y + 0z = 0.0$

$0x + 0y + 1z = 0.0$

This reads: $x = 1$, $y = 0$ and $z = 0$.

We may verify that this is a solution of the original system of equations.

NOTE

A system of equations may have

- Exactly one solution
- More than one solution
- No solution.

Our method will generate one of the solutions or imply that there is no solution.

Example 4

Solve

$x + y = 2$

$–2x – 2y = 3$

Table 0

ROW	X	Y	RHS
1	1	1	2
2	–2	–2	3

We must make the element in cell (2, 1) = 0. This is done by adding (2 x row 1) to row 2.

Table 1

ROW	X	Y	RHS
1	1	1	2
2	0	0	7

We may immediately stop since row 2 of Table 1 reads

$0x + 0y = 7$. That is,

$0 = 7$!

And this is impossible. Therefore the system of equations has no solutions.

Example 5

Acme Limited is planning to manufacture two toy trucks: Mini and Jumbo. It takes two hours to produce a batch of Minis and three hours to produce a batch of Jumbos. Each batch of Minis takes 3 tons of material 1 and each batch of Jumbos takes 7 tons of material 1. Each batch of Minis takes 3 tons of material 2 and each batch of Jumbos takes 2 tons of material 2. Suppose there are 7 hours in a day and for each day, 13 tons of material 1 and 8 tons of material 2 are available. How many batches of each toy should be produced to utilize exactly all available labor and material resources?

Let x represent the number of batches of Minis that should be produced, and y the number of batches of Jumbos.

Then:

$2x + 3y = 07$ (Labor)

$3x + 7y = 13$ (Material 1)

$3x + 2y = 08$ (Material 2).

Table 1

EQUATION	X	Y	RHS
1	2.0	3.0	7.0
2	3.0	7.0	13.0
3	3.0	2.0	8.0

Table 2

EQUATION	X	Y	RHS
1	1.0	1.5	3.5
2	0.0	2.5	2.5
3	0.0	–2.5	–2.5

Table 3

EQUATION	X	Y	RHS
1	1.0	0.0	2.0
2	0.0	1.0	1.0
3	0.0	0.0	0.0

The solution can now be read off Table 3.

$x = 2$ and $y = 1$.

That is, each toy try should be produced thus: 2 batches of Minis and 1 batch of Jumbos.

Till now we have been dealing with linear equations. Now let us go further and look at a method which helps us to find the roots of a quadratic equation.

SOLVING A QUADRATIC EQUATION

In polynomials you have seen expressions of the form $x^2 + 3x - 4$. Also we know that when an expression is equated to zero or some other expression, we call it an equation. The equations of the second degree in a single variable "x" or "y" are generally referred to as quadratic equations and the most general form of it is $ax^2 + bx + c = 0$. The roots or solution for the quadratic equation can be obtained by substituting different values for x and selecting that value for which the value of the equation is zero. The methods which we have seen in factorization of polynomials are also applicable to obtain the roots of a quadratic equation. However, in this part we look at a specific method which is only applicable to solve the quadratic equations.

According to this method, the roots of a quadratic equation $ax^2 + bx + c = 0$ are

$$x = \frac{-b-\sqrt{b^2-4ac}}{2a} \text{ and } x = \frac{-b+\sqrt{b^2-4ac}}{2a}$$

This is derived as follows. We have

$$ax^2 + bx + c = 0$$

$$ax^2 + bx = -c \quad \text{.....(1)}$$

On dividing equation (1) by a, we have $x^2 + \frac{b}{a}x = \frac{-c}{a}$

In order to make the LHS a perfect square, we add $\frac{b^2}{4a^2}$ to the LHS and since the equality is to be preserved we do so for the other side also. Hence we obtain

$$x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} = \frac{b^2}{4a^2} - \frac{c}{a}$$

That is, $\left(x + \frac{b}{2a}\right)^2 = \frac{b^2-4ac}{4a^2}$

$$x + \frac{b}{2a} = \frac{\pm\sqrt{b^2-4ac}}{2a}$$

$$x = \frac{-b}{2a} \pm \frac{\sqrt{b^2-4ac}}{2a}$$

$$x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

The following example should make this concept clear.

Example 6

Find the roots of the equation $3x^2 + 10x - 32 = 0$.

In the given quadratic equation a = 3, b = 10 and c = – 32. We now substitute these values in the formula.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$x = \frac{-10 \pm \sqrt{(10)^2 - 4(3)(-32)}}{2(3)}$$

$$x = \frac{-10 \pm \sqrt{484}}{6}$$

$$x = \frac{-10 + 22}{6} \text{ or } \frac{-10 - 22}{6}$$

$$x = 2 \text{ or } \frac{-16}{3}$$

That is, the roots of the equation $3x^2 + 10x - 32 = 0$ are x = 2 or x = –16/3. In other words, the factors are x – 2 = 0 and x + (16/3) = 0.

SUMMARY

- The Statement which equates two algebraic expressions is called an equation. In other words, the left side of the equation should be equal to its right side i.e., LHS = RHS. This can be achieved by applying 4 methods wherever required.
- The equations of second degree in a single variable are called “quadratic equation”.

Lesson 4

Indices

After reading this lesson, you will be conversant with:

- Theory of Indices
- Laws of Indices

THEORY OF INDICES

In algebra, knowing that $2^3 = 8$ is not sufficient. Equally important to know is what would be the result if quantities like $2^3 . 2^{-4} . 2^6$ or $3^7 / 3^2$ are simplified. Mind you, finding the value of quantities like these in most of the problems is not an end in itself. The values of these quantities form an input for solving the problem further. Hence, simplifying these quantities help us to solve more advanced problems. Also one feels monotonous if he tries to simplify quantities like these by stating at each step what they literally mean. In this part, we learn about the laws of indices and understand the logic behind these concepts.

LAWS OF INDICES

Law 1

$a^m \times a^n = a^{m+n}$, when m and n are positive integers.

By the above definition, a^m = a x a ..... to m factors and

a^n = a x a .... to n factors.

$a^m \times a^n$ = (a x a...to m factors) (a x a...to n factors)

= a x a ..... to m + n factors

= a^{m+n}

Now we extend this logic to negative integers and fractions. First, let us consider this for negative integer, that is, m will be replaced by – n. By the definition of $a^m \times a^n = a^{m+n}$, we get

$$a^{-n} \times a^n = a^{-n+n} = a^0$$

But we know that $a^0 = 1$(this will be proved shortly).

Hence, $a^{-n} = \frac{1}{a^n}$ or $a^n = \frac{1}{a^{-n}}$.

Similarly, what would be the case if m = p/q and n = p/q. By definition, we have

$$a^{p/q} \times a^{p/q} = a^{p/q + p/q} = a^{2p/q}$$

This can be written as $\left(a^{p/q}\right)^2 = \sqrt[q]{a^{2p}}$. This is similar to taking the qth root of a^{2p}.

Now what would be the result if we proceed to multiply $a^{p/q}$, q number of times? That is,

$a^{p/q} \times a^{p/q} \times a^{p/q} \times a^{p/q}$ ..... to q factors will be equal to $a^{qp/q}$

We express this as $(a^{p/q})^q = a^p$, that is taking the qth root of a^p.

Apart from these we look at the meaning of a^0. In this case the value of m = 0. Therefore, by definition

$$a^0 \times a^n = a^{0+n} = a^n$$

This can be also expressed as $a^0 = \frac{a^n}{a^n} = 1$.

Now we take a numerical and check the validity of this law.

$2^6 \times 2^7$ = (2 x 2 .... to 6 factors)
(2 x 2 ..... to 7 factors)

or, 2^{6+7} = 2 x 2 ..... to (6 + 7) factors

= 2^{13} = 8192

or else,

$2^6 \times 2^7$ = (2 x 2 x 2 x 2 x 2 x 2) x
(2 x 2 x 2 x 2 x 2 x 2 x 2)

= (64)(128)

= 8192

(**Note:** The same logic can be extended to more than two factors also.)

Example 1

Simplify and find the value of $2^3 \times 3^4 \times 2^2$.

We write the given quantity as $2^3 \times 2^2 \times 3^4$

$= 2^{3+2} \times 3^4$

$= 2^5 \times 3^4 = 32 \times 81 = 2592.$

Example 2

Simplify and find the value of $2x^{1/2} \cdot 3x^{-1}$, if x = 4.

We have, $2x^{1/2} \cdot 3x^{-1}$

$= 6x^{1/2} \cdot x^{-1}$

$= 6x^{1/2-1} = 6x^{-1/2}$

$= \frac{6}{x^{1/2}} = \frac{6}{4^{1/2}} = \frac{6}{(2^2)^{1/2}} = \frac{6}{2} = 3.$

Example 3

Simplify $6ab^2c^3 \times 4b^{-2}c^{-3}d$.

We have, $6ab^2c^3 \times 4b^{-2}c^{-3}d$

$= 24 \times a \times b^2 \times b^{-2} \times c^3 \times c^{-3} \times d$

$= 24 \times a \times b^{2+(-2)} \times c^{3+(-3)} \times d$

$= 24 \times a \times b^{2-2} \times c^{3-3} \times d$

$= 24ab^0 \times c^0 \times d$

$= 24ad.$

Law 2

$a^m \div a^n = a^{m-n}$, when m and n are positive integers and m > n.

By definition, a^m = a x a ..... to m factors and

a^n = a x a ..... to n factors

Therefore, $a^m \div a^n = \frac{a^m}{a^n} = \frac{\text{a x a ..... to m factors}}{\text{a x a ..... to n factors}}$

= a x a ..... to m – n factors

$= a^{m-n}$

Now, we take a numerical and check the validity of this law.

$2^7 \div 2^4 = \frac{2^7}{2^4} = \frac{\text{2 x 2 .....to 7 factors}}{\text{2 x 2 .....to 4 factors}}$

$= 2 \times 2 \times 2$.....to (7 – 4) factors

$= 2 \times 2 \times 2$.....to 3 factors

$= 2^3 = 8$

or else,

$2^7 \div 2^4 = \frac{2^7}{2^4} = \frac{\cancel{2} \times \cancel{2} \times \cancel{2} \times \cancel{2} \times \cancel{2} \times 2 \times 2 \times 2}{\cancel{2} \times \cancel{2} \times \cancel{2} \times \cancel{2}}$

$= 2 \times 2 \times 2 = 2^{1+1+1} = 2^3$

$= 8$

Example 4

Simplify $\frac{4x^{-1}}{x^{-1/3}}$.

We have, $\frac{4x^{-1}}{x^{-1/3}}$

$= 4x^{-1-(-1/3)} = 4x^{-1+1/3} = 4x^{-2/3}$ or $\frac{4}{x^{2/3}}$.

Example 5

Simplify $\frac{2a^{1/2} \text{ x } a^{2/3} \text{ x } 6a^{-7/3}}{9a^{-5/3} \text{ x } a^{3/2}}$, if a = 4.

We have, $\frac{2a^{1/2} \text{ x } a^{2/3} \text{ x } 3.2a^{-7/3}}{3.3a^{-5/3} \text{ x } a^{3/2}}$

$= \frac{2a^{1/2} \text{ x } a^{2/3} \text{ x } 2.a^{-7/3}}{3a^{-5/3+3/2}} = \frac{4a^{1/2+2/3-7/3}}{3a^{\frac{-10+9}{6}}}$

$= \frac{4a^{\frac{3+4-14}{6}}}{3a^{-1/6}}$

$= \frac{4a^{-7/6}}{3a^{-1/6}}$

$= \frac{4}{3} \text{ x } a^{-7/6-(-1/6)}$

$= \frac{4}{3} \text{ x } a^{-7/6+1/6} = \frac{4}{3} \text{ x } a^{\frac{-7+1}{6}}$

$= \frac{4}{3} \text{ x } a^{-6/6} = \frac{4}{3} \text{ x } a^{-1}$

$= \frac{4}{3a} = \frac{4}{3 \text{ x } 4} = \frac{1}{3}$.

Law 3

$(a^m)^n = a^{mn}$, when m and n are positive integers.

By definition, $(a^m)^n$ = $a^m \text{ x } a^m \text{ x } a^m$ .... to n factors.

= (a x a ..... to m factors).

(a x a ..... to m factors).

(a x a ... to m factors) ..... to n times

= a x a ..... to mn factors

= a^{mn}

Now let us look whether this is true for positive fractions. We will keep m as it is and replace n by p/q, where p and q are positive integers. Then we will have

$(a^m)^n = (a^m)^{p/q}$

Now the qth power of $(a^m)^{p/q}$ = $\{(a^m)^{p/q}\}^q$

$$= (a^m)^{\frac{p}{q} \times q}$$

$$= (a^m)^p$$

$$= a^{mp}$$

If we take the qth root of the above, we obtain

$$(a^m)^{p/q} = \sqrt[q]{a^{mp}}$$

For n being any negative quantity: In this case also m remains the same and n be replaced by – r, where r is positive. Then we have,

$$(a^m)^n = (a^m)^{-r} = \frac{1}{(a^m)^r}$$

$$= \frac{1}{a^{mr}} = a^{-mr}$$

Again replacing –r by n, we obtain a^{mn}.

Now with the help of a numerical, let us verify this law.

$$(2^4)^3 = 2^4 \times 2^4 \times 2^4$$

$$= 2^{4+4+4}$$

$$= 2^{12} = 4096$$

or else,

$$(2^4)^3 = (2^4)\,(2^4)\,(2^4)$$

$$= (2 \times 2 \times 2 \times 2)\,(2 \times 2 \times 2 \times 2)$$

$$(2 \times 2 \times 2 \times 2)$$

$$= (16)\,(16)\,(16)$$

$$= 4096.$$

Example 6

Simplify $(x^{1/2}\, y^{-1/2})^{4/3} \div (x^2\, y^{-1})^{-1/3}$

We have, $\dfrac{(x^{1/2} \,.\, y^{-1/2})^{4/3}}{(x^2 \,.\, y^{-1})^{-1/3}}$

$$= \frac{(x^{1/2})^{4/3} \,.\, (y^{-1/2})^{4/3}}{(x^2)^{-1/3} \,.\, (y^{-1})^{-1/3}}$$

$$= \frac{x^{1/2 \times 4/3} \,.\, y^{-1/2 \times 4/3}}{x^{2 \times -1/3} \,.\, y^{-1 \times -1/3}} = \frac{x^{2/3} \,.\, y^{-2/3}}{x^{-2/3} \,.\, y^{+1/3}}$$

$$= x^{2/3 - (-2/3)} \,.\, y^{-2/3 - (+1/3)}$$

$$= x^{(2/3) + (2/3)} \,.\, y^{(-2/3) - (1/3)}$$

$$= x^{4/3} \,.\, y^{-3/3}$$

$$= x^{4/3} \,.\, y^{-1} = \frac{x^{4/3}}{y} \,.$$

Example 7

Simplify $\sqrt[3]{x^{-1}\sqrt{y^3}} \div \sqrt[3]{y \,.\, \sqrt[3]{x}}$

We have, $\sqrt[3]{x^{-1}\sqrt{y^3}} \div \sqrt[3]{y.\sqrt[3]{x}}$

$$= \left(x^{-1}.\sqrt{y^3}\right)^{1/3} \Big/ \left(y.\sqrt[3]{x}\right)^{1/3}$$

$$= (x^{-1})^{1/3}.(y^{3/2})^{1/3} \Big/ (y)^{1/3}.(x^{1/3})^{1/3}$$

$$= \frac{x^{-1/3}.y^{3/2 \times 1/3}}{y^{1/3}.x^{1/3 \times 1/3}} = \frac{x^{-1/3}.y^{1/2}}{y^{1/3}.x^{1/9}}$$

$$= x^{-1/3-1/9}.y^{1/2-1/3}$$

$$= x^{-4/9}.y^{1/6}$$

Law 4

$(ab)^n = a^n b^n$, where n can take all of the values.

First, we look at n when it is a positive integer. Then by definition, we have

$(ab)^n$ = ab x ab ..... to n factors

= (a x a ..... to n factors)
(b x b ..... to n factors)

= a^n x b^n

When n is a positive fraction, we will replace n by p/q.

Then we will have $(ab)^n = (ab)^{p/q}$

The qth power of $(ab)^{p/q} = \{(ab)^{p/q}\}^q$

$= (ab)^p$

$= a^p$ x b^p

$= (a^{p/q}$ x $b^{p/q})^q$

On taking the qth root of this, we obtain

$(ab)^{p/q} = a^{p/q}$ x $b^{p/q}$

When n takes any negative value: That is we replace n by – r, where r is a positive number. Then

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

$$= \frac{1}{(a^r b^r)}$$

$$= a^{-r} \text{ x } b^{-r}$$

If we replace – r by n, we have $(ab) = a^n$ x b^n x n.

Example 8

Simplify $(x^a y^{-b})^3 . (x^3 y^2)^{-a}$.

We have $(x^a y^{-b})^3 . (x^3 y^2)^{-a}$

$= (x^a)^3 . (y^{-b})^3 . (x^3)^{-a} . (y^2)^{-a}$

$= x^{3a} . y^{-3b} . x^{-3a} . y^{-2a}$

$= x^{3a+(-3a)} . y^{-3b+(-2a)}$

$= x^0 . y^{-3b-2a}$

$= 1 . y^{-2a-3b}$

$= y^{-2a-3b}$ or $1/y^{2a+3b}$

Example 9

Simplify $\sqrt[6]{a^{4b}\, x^6} \cdot (a^{2/3}\, x^{-1})^{-b}$

We have, $\sqrt[6]{a^{4b}\, x^6} \cdot (a^{2/3}\, x^{-1})^{-b}$

$= (a^{4b}\, x^6)^{1/6} \cdot (a^{2/3})^{-b} \cdot (x^{-1})^{-b}$

$= (a^{4b})^{1/6}\, (x^6)^{1/6} \cdot a^{2/3\, x\, -b} \cdot x^{-1\, x\, -b}$

$= a^{4b/6} \cdot x^{6\, x\, 1/6} \cdot a^{-2b/3} \cdot x^{b}$

$= a^{4b/6\, -\, 2b/3} \cdot x^{1\, +\, b}$

$= a^{0} \cdot x^{1+b} = x^{1+b}$

SUMMARY

- In an expression 3^2. 2 is called a power or index or exponent.
- The lesson specifically provides an insight into various laws of indices, which explain the basic idea of the concept.

Lesson 5

Progressions

After reading this lesson, you will be conversant with:

- Arithmetic Progression
- Geometric Progression
- Harmonic Progression

In this section, we will look at three types of progressions called Arithmetic, Geometric and Harmonic Progression. Before we start looking at the intricacies of these let us understand what is meant by series. A series is a collection of numbers which may or may not terminate at some point. The first set of series is called finite series and the second one infinite series. In the theoretical sense, an infinite series conveys that the number of elements in the series are so large that it is practically uncountable. Generally, series are expressed in an abridged form in terms of a general term known as nth term. Therefore, given a series we can obtain its nth term or else given an n th term we can obtain the different elements of that series. For example, consider a simple nth term which is:

$$T_n = \frac{(n+1)(n+2)}{2}$$

If we substitute n = 1, the value of $T_{n=1}$ will be

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

If we substitute n = 2, the value of $T_{n=2}$ will be

$$\frac{(2+1)(2+2)}{2} = 6$$

If we continue to substitute different values for n, as we did above, we get different values of this particular series. This is an example of infinite series, whereas a series like 1, 2, 3, 4, 5, 6 is an example of finite series. The general term is given by $T_n = n + 1$, where n takes values from 0 to 5. After looking at these two examples we find that a series is finite or infinite depending on the values taken by n. In other words, a series terminates depending on the extent of values taken by n. Now let us look at three types of progressions.

ARITHMETIC PROGRESSION (A.P.)

A series is said to be in Arithmetic Progression (A.P.) if the consecutive numbers in the series differs by a constant value. This constant value is referred to as "common difference". The series in which the consecutive terms increase by a constant quantity, is referred to as an increasing series and if the terms decrease by a constant quantity it is referred to as a decreasing series. The series

3, 7, 11, 15, 19, ...

is an example of increasing series, while the one like

8, 2, –4, ...

is an example of decreasing series.

In an A.P. the first number is denoted by "a" and the common difference is denoted by "d". If we know the values of a and d, it is quite easy to get the terms of the Arithmetic Progression. In terms of a and d, the consecutive terms of arithmetic progression are

a, a + d, a + 2d, a + 3d, ... a + nd

We observe that the first term is a, the second term is a + d, the third term being a + 2d. The point to note is that for the first term the coefficient of d is zero, for the second term it is one and for the third term it is 2. By observing this pattern can we conclude that the coefficient of nth term is n – 1? Yes, we can. In fact, the nth term is given by

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

Generally the T_n term is also denoted by "l" (small alphabet 'l'). That is, $l = a + (n - 1)d$.

Now let us look at an example.

Example 1

If the first term of an A.P. 'a' = 3 and the common difference 'd' = 2, what are the first five terms of the series and what would be the nth term? They are calculated as follows. We know that

$T_1 = a = 3$

$T_2 = a + d = 3 + 2 = 5$

$T_3 = a + 2d = 3 + 2(2) = 7$

$T_4 = a + 3d = 3 + 3(2) = 9$

$T_5 = a + 4d = 3 + 4(2) = 11$

: :

: :

$l = T_n = a + (n - 1)d = 3 + (n - 1)(2)$

$= 3 + 2n - 2$

$= 2n + 1.$

Sum of a Number of Terms in A.P.

We know that the terms in an A.P. are given by

a, a + d, a + 2d, a + 3d, ... a + (n – 2)d, a + (n – 1)d

The sum of all these terms which is denoted by "s" is given by

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

This is obtained as follows. We know that

$s = (a) + (a + d) + (a + 2d) + (a + 3d) + ... +$

$(a + (n - 2)d) + (a + (n - 1)d)$

Now we reverse the order and write it as shown below.

$s = (a + (n - 1)\, d) + (a + (n - 2)\, d) + ... +$

$(3d + a) + (2d + a) + (d + a) + a$

On adding the respective terms we get

$2s = \{a + a + (n - 1)d\} + \{a + d + a + (n - 2)d\}$

$+ ..... + \{a + (n - 2)d + a + d\} +$

$\{a + (n - 1)d + a\}$

That is, we have:

$2s = \{2a + (n - 1)d\} + \{2a + d + (n - 2)d\} +$

$..... + \{2a + d + (n - 2)d\} +$

$\{2a + (n - 1)d\}$

Further simplifying we obtain

$2s = \{2a + (n - 1)d\} + \{2a + d + nd - 2d\} + ... +$

$\{2a + d + nd - 2d\} + \{2a + (n - 1)d\}$

This would be

$2s = \{2a + (n - 1)d\} + \{2a + d + nd - 2d\} + ... +$

$\{2a + d + nd - 2d\} + \{2a + (n - 1)d\}$

On simplification we obtain

$2s = \{2a + (n-1)d\} + \{2a + nd - d\} + \ldots\ldots\ldots +$
$\{2a + nd - d\} + \{2a + (n-1)d\}$

$2s = \{2a + (n-1)d\} + \{2a + (n-1)d\} + \ldots\ldots\ldots +$
$\{2a + (n-1)d\} + \{2a + (n-1)d\}$

Since 2 x 2 x 2 x 2 = 2(1 + 1 + 1 + 1) = 2 x 4,

$2a + (n-1)d$ multiplied n times will be $n.\{2a + (n-1)d\}$. Therefore,

$$2s = n.\{2a + (n-1)d\}$$

or

$$s = \frac{n}{2}\{2a + (n-1)d\} \quad \ldots\ldots\ldots\ldots (a)$$

Since $l = a + (n-1)d$, equation (a) is also written as

$$s = \frac{n}{2}\{a + a + (n-1)d\} \text{ or}$$

$$s = \frac{n}{2}\{a + l\}$$

Now we will find the sum of 20 terms when a = 5 and d = 2. Substituting these values in the formula, we obtain

$$s = \frac{20}{2}\{2(5) + (20-1)2\}$$
$$= 480$$

This problem can also be solved by finding the last term which in this case happens to be T_{20} and it is given by $T_{20} = 5 + (20-1)2 = 43$. Therefore,

$$s = \frac{n}{2}\{a + l\}$$
$$= \frac{20}{2}\{5 + 43\} = 480.$$

We observe that both these methods are essentially the same. With this background let us look at few more examples.

Example 2

For the series given below, find the 23rd and the 27th terms.

38, 36, 34, ...

We find the first term that is a = 38. The common difference d is given by 36 – 38 = –2. The 23rd term is given by

$T_{23} = a + 22d$
$= 38 + 22(-2)$
$= 38 - 44 = -6$

Similarly the 27th term is given by

$T_{27} = a + 26d$
$= 38 + 26(-2)$
$= 38 - 52$
$= -14$

Example 3

Given that the first term of a series is 7 and the last term of the same to be 49. The sum of the series is 336. Find the number of elements in this series and the common difference.

We know that the sum of n terms is given by

$$s = \frac{n}{2}\{a + l\}$$

$$336 = \frac{n}{2}\{7 + 49\}$$

$$672 = 56n$$

$$n = 672/56$$

$$= 12$$

That is, the number of terms is n = 12, and this term happens to be 49. That is,

$$T_{12} = a + 11d$$

$$49 = 7 + 11d$$

$$49 - 7 = 11d$$

$$\text{or } d = 42/11 = 3\frac{9}{11}$$

Arithmetic Mean

When three quantities are in A.P., then the middle one is said to be the arithmetic mean of the other two. That is, if a, b and c are in A.P., then b is the arithmetic mean of a and c. Since a, b and c are in A.P., the common difference ought to be constant, we have

$$b - a = c - b$$

$$\text{or } b + b = a + c$$

$$\text{or } 2b = a + c$$

$$\text{or } b = \frac{a + c}{2}$$

We should remember that between any two quantities we can insert any number of terms so that the resultant series is in A.P. Now we look at some examples wherein we will insert the required number of terms.

Example 4

Insert 15 arithmetic means between 4 and 68.

We are given the first term and the last term. The total number of terms including 4 and 68 are therefore 17. Since "a" is given we ought to find "d". The 17th term is given by $T_{17} = a + 16d = 4 + 16d$. Also the T_{17} term is given to be 68. Therefore,

$$68 = 4 + 16d$$

$$16d = 68 - 4$$

$$16d = 64$$

$$d = 64/16 = 4$$

Using the values of "a" and "d" we insert the required terms. The series will be 4, 8, 12, ..., 68.

Example 5

The sum of three terms in an A.P., is 39 and their product is 2184. Find the terms.

If we assume "a" to be the middle term of this series, then the three terms could be a – d, a and a + d. The sum of these three terms is

$$a - d + a + a + d = 3a$$

and this is equal to 39. Therefore, 3a = 39, which gives the value of a = 13. The product of these three terms is given by (a – d)(a)(a + d) and its value is given by 2184. That is,

$$(a - d)(a)(a + d) = 2184$$

$$(13 - d)(13)(13 + d) = 2184$$

$$169 - d^2 = 2184/13 = 168$$

[We employ the identity $(a + b)(a – b) = a^2 – b^2$]

$$d^2 = 169 - 168 = 1$$

$$d = \sqrt{1} = \pm 1$$

Considering only the positive value of d, the three terms are 13 – 1, 13 and 13 + 1. That is, 12, 13 and 14 respectively.

(**Note:** In this case even if we consider d = –1, the three terms we obtain are 13 – (–1), 13 and 13 + (–1), which are 14, 13 and 12 respectively. The only difference is that this happens to be decreasing series.)

Example 6

The sum of five numbers in an A.P. is 40, and the sum of their squares is 410. Find the numbers.

Let us assume the numbers to be a – 2d, a – d, a, a + d and a + 2d. The sum of these numbers is given to be 40. That is,

$$(a - 2d) + (a - d) + a + (a + d) + (a + 2d) = 40$$

$$5a = 40$$

$$a = 40/5 = 8$$

Also given that the sum of the squares of these terms as 410. That is,

$$(8 - 2d)^2 + (8 - d)^2 + (8)^2 + (8 + d)^2 + (8 + 2d)^2 = 410$$

Expanding these we obtain

$$4d^2 - 32d + 64 + d^2 - 16d + 64 + 64 + d^2 - 16d + 64 + 4d^2 - 32d + 64 = 410$$

(For expansion, we employed $(a + b)^2 = a^2 + 2.a.b + b^2$ and $(a – b)^2 = a^2 – 2.a.b + b^2$)

By canceling out and rearranging the terms we have

$$10d^2 + 320 = 410$$

$$10d^2 = 410 - 320$$

$$d^2 = 90/10 = 9$$

That is, $d = \pm 3$

We consider only the positive value of d and get the individual terms. They are:

8 – 2(3), 8 – (3), 8, 8 + 3 and 8 + 2(3)

Simplifying these we obtain 2, 5, 8, 11 and 14.

GEOMETRIC PROGRESSION (G.P.)

Learning geometric progression in comparison with arithmetic progression is easier. In geometric progression also, we denote the first term by 'a' but a basic difference from A.P. is that instead of common difference we have common ratio

'r'. Like d, r remains constant whenever the ratio of any two consecutive terms is computed. The terms of a G.P. are:

$$a, ar, ar^2, ar^3, ar^4, ..., ar^{n-1}$$

That is, $T_1 = a$

$T_2 = ar$

$T_3 = ar^2$

: :

: :

$T_n = ar^{n-1}$

This we observe is similar to A.P. We take an example to become more familiar with this.

Example 7

It is known that the first term in a G.P. is 3 and the common ratio r is 2. Find the first three terms of this series and also the nth term.

We know that the first term is given by

$T_1 = a = 3$

$T_2 = ar = 3.2 = 6$

$T_3 = ar^2 = 3.2.2 = 12$

The nth term is given by $T_n = ar^{n-1} = 3(2)^{n-1}$.

Sum of a Number of Terms in a G.P.

We know that the terms in a G.P. are:

$$a, ar, ar^2, ar^3, ar^4, ..., ar^{n-1}$$

Let s be the sum of these terms, then

$$s = a + ar + ar^2 + ar^3 + ar^4 + ...+ ar^{n-1}$$

or

$$s = \frac{a(1-r^n)}{(1-r)}$$

This is obtained as follows:

We know that

$$s = a + ar + ar^2 + ar^3 + ar^4 +.....+ ar^{n-1} \quad .....(1)$$

Multiplying this with "r" throughout, we have

$$r.s = r.a + r.ar + r.ar^2 + r.ar^3 + r.ar^4 +.....+ r.ar^{n-1}$$

$$= ar + ar^2 + ar^3 + ar^4 + ar^5 +.....+ ar^n \quad .....(2)$$

Subtracting (1) from (2), we have

$$r.s - s = (ar - a) + (ar^2 - ar) + (ar^3 - ar^2) +.....+ (ar^n - ar^{n-1})$$

After canceling the terms equal in magnitude but opposite in sign, we are left with

$s(r-1) = ar^n - a$

$s(r-1) = a(r^n - 1)$

or $s = \frac{a(r^n-1)}{(r-1)}$

By changing the signs in the numerator and the denominator we can also write the above equation as

$$s = \frac{a(1-r^n)}{(1-r)}$$

What happens to the above formula if the value of n is very large? The above formula can be written as

$$s = \frac{a}{(1-r)} - \frac{ar^n}{(1-r)}$$

As the value of n approaches infinity (very large) the expression $\frac{ar^n}{(1-r)}$ becomes smaller to that extent where we ignore it. In this case, the nth term is given as

$$T_n = \frac{a}{(1-r)}$$

Now we look at a couple of examples.

Example 8

Find the sum of the series which is given below to 13 terms.

81, 54, 36, ...

The first term 'a' = 81 and the common ratio is obtained from the ratio of 54 and 81 or 36 and 54. It is 54/81 = 2/3. Now we employ the formula given above to calculate the sum of series to 13 terms.

$$s = \frac{a(1-r^n)}{(1-r)} = \frac{(81)(1-(2/3)^{13})}{(1-2/3)} = \frac{(81)(0.995)}{1/3}$$

$$= 241.78$$

The same series if considered as an infinite series, the sum of n terms would be

$$T = \frac{a}{(1-r)} = \frac{81}{1-2/3} = 243$$

Geometric Mean

When three quantities a, b and c are in G.P., then the geometric mean "b" is calculated as follows.

Since these quantities are in G.P., the ratio of b/a and the ratio of c/b should give us the same number. In other words, these ratios should be equal. That is,

$$\frac{b}{a} = \frac{c}{b}$$

On cross multiplying these, we have b.b = a.c. That is, $b^2 = ac$.

Example 9

Find the geometric mean between the two numbers 36 and 40.

If c is the number between these two numbers, then the geometric mean is given by

$$c^2 = (36)(40)$$

$$c = \sqrt{1440} = 37.95$$

Now with the help of examples we look at how to insert the required number of terms in between two given quantities.

Example 10

Insert 6 geometric means between 5 and 640.

Taking into account the first and the last terms 5 and 640, we have 8 terms in all. We are given the first and the eighth terms. The first term 'a' = 5 and the eighth term

$T_8 = ar^7 = 5.r^7$

But $T_8 = 640 = 5.r^7$

$r^7 = \frac{640}{5} = 128$

$r = (128)^{1/7} = (2^7)^{1/7} = 2$

Employing a = 5 and r = 2, the six geometric means are 10, 20, 40, 80, 160 and 320.

HARMONIC PROGRESSION

Three quantities a, b and c are said to be in Harmonic progression if

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

In this case, we observe that we have to consider three terms in order to conclude whether they are in Harmonic progression or not.

An important proposition in this case is that the reciprocal of quantities in Harmonical Progression are in Arithmetical Progression. Let us understand this by considering three quantities a, b and c. By definition, if a, b and c are in Harmonic Progression, then they satisfy the condition that

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

By cross multiplying, we obtain

$a(b - c) = c(a - b)$

That is, $ab - ac = ac - bc$

Dividing each of these terms by abc, we have

$$\frac{ab-ac}{abc} = \frac{ac-bc}{abc}$$

This can be written as

$$\frac{ab}{abc} - \frac{ac}{abc} = \frac{ac}{abc} - \frac{bc}{abc}$$

Canceling the common terms, we have

$$\frac{1}{c} - \frac{1}{b} = \frac{1}{b} - \frac{1}{a}$$

This gives us the common difference between the reciprocal terms of a, b and c. This also proves our proposition.

Now we look at an example.

Example 11

If the 12th and the 19th term in an Harmonic progression are 1/5 and 3/22 respectively, find the series.

We solve this problem by taking the reciprocal of the given terms. That is, we are dealing with an A.P. The reciprocal terms are 5 and 22/3. As these are the 12th and 19th terms we have

$T_{12} = a + 11d$

$5 = a + 11d$ .....(1)

and

T_{19} = a + 18d

22/3 = a + 18d .....(2)

Subtracting (1) from (2), we have

$$a + 18d - (a + 11d) = \frac{22}{3} - 5$$

Simplifying we have

$$a + 18d - a - 11d = \frac{22}{3} - 5$$

That is, 7d = 7/3

d = 1/3

Substituting this value of d in equ. (1), we have

5 = a + 11(1/3)

a = 4/3

Now employing the values of a and d, the terms of the series are

$$\frac{4}{3}, \frac{5}{3}, 2, \frac{7}{3}, \text{.....}$$

The reciprocals of these terms are $\frac{3}{4}, \frac{3}{5}, \frac{1}{2}$, and $\frac{3}{7}$.

Harmonic Mean

If a, b and c are in harmonic progression with b as their harmonic mean then

$$b = \frac{2ac}{a+c}$$

This is obtained as follows. Since a, b and c are in Harmonic progression, 1/a, 1/b and 1/c are in arithmetic progression. Then

$$\frac{1}{b} - \frac{1}{a} = \frac{1}{c} - \frac{1}{b}$$

This can be written as

$$\frac{1}{b} + \frac{1}{b} = \frac{1}{c} + \frac{1}{a}$$

$$\frac{2}{b} = \frac{1}{c} + \frac{1}{a} = \frac{a+c}{ac}$$

On cross multiplication we obtain

2ac = b(a + c)

That is, $b = \frac{2ac}{a+c}$

The second proposition we are going to look at in this part is: If A, G and H are the arithmetic, geometric and harmonic means respectively between two given quantities a and b then $G^2 = AH$. The explanation is given below.

We know that the arithmetic mean of a and b is $\frac{a+b}{2}$ and it is given that this equals A. Similarly, $G^2 = ab$ and $H = \frac{2ab}{a+b}$.

The product of $AH = \frac{a+b}{2} \cdot \frac{2ab}{a+b} = ab$. This we observe is equal to G^2.

That is, $G^2 = AH$, which says that G is the geometric mean between A and H.

Example 12

Insert two harmonic means between 4 and 12.

We convert these numbers into A.P. They will be 1/4 and 1/12. Including the two arithmetic means we have four terms in all. We are given the first and the fourth terms. Thus

T_0 = a = 1/4 and

T_4 = a + 3d = 1/12

Substituting the value of a = 1/4 in T, we have

1/4 + 3d = 1/12

3d = 1/12 – 1/4 = – 1/6

d = –1/18

Using the values of a and d, we obtain T_2 and T_3.

T_2 = a + d = 1/4 + (–1/18)

= 1/4 – 1/18 = 7/36

T_3 = a + 2d = 1/4 + 2.(–1/18)

= 1/4 – 2/18

= 1/4 – 1/9

= 5/36

The reciprocals of these two terms are 36/7 and 36/5.

Therefore, the harmonic series after the insertion of two means will be 4, 36/7, 36/5 and 12.

SUMMARY

- A series is a collection of numbers, which may or may not terminate at some point. It can be finite or infinite in nature.
- A series can be expressed in terms of Arithmetic or Geometric or Harmonic progression.
- The “mean” forms the essence of these progressions.

Lesson 6

Permutations and Combinations

After reading this lesson, you will be conversant with:

- Permutations
- Combinations

Introduction

Consider this. You have four units A, B, C and D. You are asked to select two out of these four units. How do you go about this particular task? Will your methodology remain the same, if you are asked that you should select two units, but they should be according to some predefined criteria? Definitely, it differs. In this part we look at two techniques called Permutations and Combinations, which help us solve problems like these.

Before we start looking at permutations and combinations, let us acquaint ourselves with an important principle. It says: if an operation (first) has been performed in say 'm' ways and a second operation which can be performed in 'n' ways, then both the operations can be performed in m x n ways. The explanation is as follows.

The first operation can be performed in any one of the given m ways. After performing this operation in any one of the m ways, the second operation can be performed in any one of the n ways. Since both the operations are performed in any one of either m or n ways, why is that we get m x n ways? Here we have to understand that the first operation is performed in only one of the m ways, but with this one way, we can associate n ways of doing the second operation. In other words, we have 1 x n = n ways of performing both the operations, taking into consideration not more than one way of performing the first operation. And therefore corresponding to m ways of performing the first operation we have m x n ways of performing both the operations.

Remember that this concept can be applied even if we have more than two operations. The following example should make this concept clear.

Example 1

A person from his office can go to his residence via one of the 3 routes. In how many ways can that person go to his residence via one route and come to office by another route?

That person can go to his residence by one of the three routes. That is, he has 3 ways. Now he can come to office via one of the remaining two routes since he should not take the same route. That is, he can do so in two ways. Therefore, the number of ways that person can go to his residence and come back to his office by 3 x 2 = 6 ways.

PERMUTATIONS

Now we look at Permutations and its related concepts. Permutations are defined as each of the **arrangements** that can be made by taking some or all of the elements given. Here the word arrangement should be understood properly. This will be clear if we consider the given example of taking two out of four units elements viz., A, B, C, D. The permutations of taking two units out of four can be done in the following ways.

AB, AC, AD, BC, BD, CD

BA, CA, DA, CB, DB, DC

Here we are looking at arranging two units in a particular order. In other words, the arrangement AB is not the same as the arrangement BA and therefore, it is necessary to list both of them. Thus AB and BA both are different arrangements of two units A and B.

Finding the number of Permutations of 'n' dissimilar things taken 'r' at a time:

After looking at the definition of permutations, we look at how to evolve a general framework for finding the number of permutations of 'n' dissimilar things taken 'r' at a time. To make this simpler we again go back to our introduction example but with a slight change. Here we consider five units and one has to take four out of five units. Now in how many ways can one take the first unit? Five ways. Since he can take any one of the five units. After taking the first unit in five ways, in how many ways can he take the next unit? Necessarily in four ways. What about the third and the fourth units? He can take them in three and two ways respectively. At this point it is easy to observe the pattern.

Now applying the principle we have studied above can we state that four units out of five can be taken in

5 x 4 x 3 x 2 ways

Yes, we can and this principle forms the basis for finding the number of permutations of n dissimilar things taken r at a time. Therefore, the first thing can be taken in n ways, the second thing can be taken in n – 1 ways, the third thing can be taken in n – 2 ways and in a similar fashion the r^{th} thing can be taken in n – (r – 1) ways. Why n – (r – 1)? Because the first thing is taken in n – (1 – 1) = n ways, the second in n – (2 – 1) = n – 1 ways, ... the rth thing in n – (r – 1) = n – r + 1 ways. From the principle, taking r dissimilar things from n things is therefore

n.(n – 1).(n – 2).....(n – r +1) ways.

What we will get if we have to take all the given things at a time. It will be n. (n – 1).(n – 2).(n – 3).....to n factors. That is,

n.(n – 1).(n – 2).(n – 3).....3.2.1

But this happens to be the definition of n factorial, denoted as n! At this stage only remember that n! = n.(n – 1).(n – 2).(n – 3).....3.2.1

Taking r things out of n things is denoted by ${}^{n}P_{r}$ and it stands for $\frac{n!}{(n-r)!}$. On simplifying this, we get

n.(n – 1).(n – 2).....(n – r + 1)

which is taking r things out of n.

(**Note:** Since selecting r elements from n elements is similar to filling up n positions with r things, we often use this analogy in understanding concepts in Permutations and Combinations.)

Now we take up an example.

Example 2

There are six boxes and three balls. Find in how many ways can these three balls put into these six boxes?

The first ball can be put into any one of the six boxes. That is six ways. The second ball can be then put into any one of the remaining five boxes. That is in five ways and finally the last ball can be put into one of the remaining four boxes, which gives us 4 ways. That is, the three balls can be put into six boxes in

6 x 5 x 4 = 120 ways.

For the same problem, let us apply the formula and check whether we get the same answer.

$${}^{n}P_{r} = \frac{n!}{(n-r)!} = \frac{6!}{(6-3)!} = \frac{6.5.4.3!}{3!} = 120$$

Finding the number of permutations of n things taken r at a time, given that each of the elements can be repeated once, twice .... up to r times.

In this case the first place can be filled up by any one of the n values. The second position can also be filled up by any one of the n values. Similarly the third, fourth and the rth positions. This is because we have the discretion to use each element for as many as r times. Therefore, r things out of n things can be selected in n^r ways.

Example 3

Find the number of ways in which three prizes can be awarded to three students, when each student is eligible for all the prizes.

The first prize can be awarded to any one of the three students. That is, it can be given in three ways. Similarly, the second and third prizes. Therefore, the three prizes can be given away in 3^3 ways, which is 27 ways.

Till now we have been looking at situations where the elements are different from each other. On some occasions we come across situations wherein some elements are of one kind, some other elements are of one kind and the rest all different. In this part we obtain a general framework which helps us to solve problems like these.

To find the number of ways in which n things may be arranged among themselves, taking all at a time, when p of the things are alike (of one kind), q of them alike but of another kind, r of them of a third kind and the rest all different:

We have a total of n things, of which p are of one kind, q are of one kind, r of one kind and the rest that is n – (p + q + r) things being distinct. If ${}^nP_r$ is the required number of permutations and then if p things are replaced by same number of distinct things from any one of the ${}^nP_r$ permutations without disturbing the position of the remaining letters, we could form p! new permutations. And if this change is carried out in each of the ${}^nP_r$ permutations, we will obtain ${}^nP_r$ x p! permutations.

If the same procedure is carried out for q and r things, the number of permutations would be ${}^nP_r$ x p! x q! x r!. Since the things are all now different, the number of arrangements that can be made among themselves is n!. That is, n! = ${}^nP_r$ x p! x q! x r!. This can be expressed as

$$ {}^nP_r = \frac{n!}{p! \times q! \times r!} $$

which is also our required equation.

Example 4

You are given a word “MANAGEMENT” and asked to compute the number of permutations that you can form taking all the letters from this word.

We observe that the given word consists of 10 letters in all. In these 10 letters, we find two letters each of M, N, A and E. The two remaining letters are G and T. By applying the above formula, the number of permutations that can be formed by taking all the letters is

$$ = \frac{10!}{2!.2!.2!.2!} $$

COMBINATIONS

Now we take up combinations and its related concepts. Combinations are defined as each of the groups or selections which can be made by taking some or all of the elements from the given elements. Let the units/elements be A, B, C, D. The combinations of taking two units out of four units are given by

AB, AC, AD, BC, BD, CD

That is, in combinations the emphasis on order is not there and one is concerned with only the number of units that ought to be selected.

Finding the number of combinations of 'n' dissimilar things taken 'r' at a time:

To obtain this relationship, we consider a set S consisting of n elements which are distinct. To specify a permutation of size 'r' chosen from these n elements, we can first select the r elements that will appear in the permutation, and then we can give the order in which the selected elements are to be arranged. The first step constitutes the selection of a combination of r elements from set S consisting of n elements, and this can be done in ${}^{n}C_{r}$ ways. The second step constitutes the arrangement of these elements. The ordering (arrangement) of these r elements can be accomplished in r! ways. Therefore, the number of permutations of n things taken r at a time, that is ${}^{n}P_{r}$, will be the product of ${}^{n}C_{r}$ and r!. That is,

${}^{n}P_{r} = {}^{n}C_{r}$. r!. This can also be expressed as

${}^{n}C_{r} = \frac{{}^{n}P_{r}}{r!}$ which gives our required relationship. Since ${}^{n}P_{r} = \frac{n!}{n-r!}$, the expression for ${}^{n}C_{r}$ will be $\frac{n!}{n-r!.\ r!}$.

Now we look at a couple of examples.

Example 5

Find the number of combinations of 50 things taking 46 at a time.

That is, we have to select 46 things out of 50 things without giving any importance to their arrangement. That will be ${}^{50}C_{46}$. But we know that

$$
\begin{aligned}
{}^{n}C_{r} &= \frac{n!}{n-r!.\ r!} \\
&= \frac{50!}{50-46!.\ 46!} \\
&= \frac{50 \text{ x } 49 \text{ x } 48 \text{ x } 47 \text{ x } 46!}{4 \text{ x } 3 \text{ x } 2 \text{ x } 1.\ 46!} \\
&= 230300.
\end{aligned}
$$

Example 6

In how many ways can a party of 4 or more be selected from 10 persons?

The party should consist of at least 4 persons. That is, it can consist of either 4 or 5 or 6 or 7 or 8 or 9 or 10 persons. Therefore, a party of at least 4 can be selected in

$$
\begin{aligned}
&{}^{10}C_{4} + {}^{10}C_{5} + {}^{10}C_{6} + {}^{10}C_{7} + {}^{10}C_{8} + {}^{10}C_{9} + {}^{10}C_{10} \\
&= \frac{10!}{10-4!.\ 4!} + \frac{10!}{10-5!.\ 5!} + \frac{10!}{10-6!.\ 6!}
\end{aligned}
$$

$$+\frac{10!}{10-7!.\ 7!}+\frac{10!}{10-8!.\ 8!}+\frac{10!}{10-9!.\ 9!}+\frac{10!}{10-10!.\ 10!}$$

$$=\frac{10!}{6!.4!}+\frac{10!}{5!.5!}+\frac{10!}{4!.6!}+\frac{10!}{3!.7!}+\frac{10!}{2!.8!}+\frac{10!}{1!.9!}+\frac{10!}{0!.10!}$$

[**Note:** 0! = 1]

On simplification we obtain

= 210 + 252 + 210 + 120 + 45 + 10 + 1

= 848.

The number of combinations of n things taken r at a time is equal to the number of combinations of n things taken n – r at a time:

Proving this statement should not be a problem. It says that

$^{n}C_{r} = {}^{n}C_{n-r}$

We know that $^{n}C_{r} = \frac{n!}{n-r!.\ r!}$. The same when applied to $^{n}C_{n-r}$ will be

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

$$^{n}C_{n-r} = \frac{n!}{n-n+r!.\ n-r!}$$

that is, $^{n}C_{n-r} = \frac{n!}{r!.\ n-r!}$

which is $^{n}C_{r}$. Combinations like these are called complementary.

To find the total number of ways in which it is possible to make a selection by taking some or all of n things:

In this case we can either consider a thing (select it) or else leave it behind (do not select it). That is, there are two ways of dealing with a particular element. Since either of these two ways can be associated with either of the two ways of dealing with another thing, the number of dealing with n things is given by

2 x 2 x 2 x 2 ..... to n factors.

This gets simplified to 2^{n}. But we also have a case wherein we may not select a single element. Considering that this can be done in a single way we subtract it from the number of ways of selecting some or all of n things. Therefore, the total number of ways of doing this is given by $2^{n} - 1$. Since we are concerned with selections only this is referred to as "total number of combinations" of n things.

Example 7

A person has 5 friends. He wants to invite one or more of them for dinner. In how many ways can he do that?

In this example, the person can select some or all of his friends to dinner. Applying the above principle he can do so in $2^{5} - 1 = 31$ ways.

Now let us look at how to solve problems where some elements in the given elements are of one kind, some others of second kind and the rest all different.

To find the total number of ways in which it is possible to make a selection by taking some or all out of p + q + r + ... things, where p of them are alike, q of them are alike (second kind), r alike of the same kind and so on:

The p things can be selected in p + 1 ways. Why p + 1 things? This also includes not selecting at least one from p things plus selecting p things in p ways. Therefore, q things can be selected in q + 1 ways, r things are selected in r + 1 ways and so on. Hence the number of ways in which all the things may be selected is given by

$(p + 1)(q + 1)(r + 1).....$

ways. Since this also includes the case wherein we do not select at least one thing, the total number of combinations is given by

$(p + 1)(q + 1)(r + 1)..... - 1$ things.

To find the number of ways in which m + n things can be divided into two groups containing m and n elements respectively:

Finding the number of ways in which m + n things can be divided into two groups containing m and n things respectively is same as finding the number of combinations of m + n things m at a time. That is, every time we select a group of m things we leave behind the group containing n things. Thus the required number of ways is given by $\frac{m+n!}{m!.n!}$

SUMMARY

- Permutations are the arrangements that can be made by some or all of the elements given.
- Combinations are groups made by taking some or all elements of the given elements.
- Permutations and combinations are the essence of any study undertaken and can be done in many ways.

Lesson 7

Logarithms

After reading this lesson, you will be conversant with:

- Rules of Logarithms
- Transforming the Base of Logarithms

Introduction

We know that $2^4 = 16$ and also that 2 is referred to as the base, 4 as the index or power or the exponent. The same if expressed in terms of logarithms would be $\log_2 16 = 4$ and is read as the logarithm of 16 to base 2 is 4. Hence we define the logarithm of a number to a given base as the index of the power to which the base should be raised in order to equal the given number. We look at the following example.

What would be the value of $\log_{12}144$?

If we assume x to be the value then

$$\log_{12}144 = x$$

This is the same as $144 = 12^x$. That is, 12 should be raised or in other words multiplied by itself so that the resultant value is 144. We find that 12 when multiplied twice would give 144. That is, the value of x = 2. This gives the value of $\log_{12}144$ as 2.

RULES OF LOGARITHMS

Rule 1

The logarithm of 1 to any base is 0.

Proof

We know that any number raised to zero equals 1. That is, $a^0 = 1$, where "a" takes any value. Therefore, the logarithm of 1 to the base a is zero. Mathematically, we express this as $\log_a 1 = 0$.

Example 1

What is the value of $\log_{10}1$?

Needless to say this would be zero.

Rule 2

The logarithm of a number, the number being same as the base is 1.

Proof

We know that any number raised to the power of 1 is itself. That is $a^1 = a$. Therefore, the logarithm of a to the base a is equal to 1.

Mathematically, we express this as $\log_a a = 1$.

Example 2

What is the value of $\log_{13}13$?

By applying the above rule, the value of $\log_{13}13$ is 1.

Rule 3

The logarithm of a product to base a is equal to sum of the logarithms of the individual numbers which constitute the product to the same base a. That is, $\log_a M.N = \log_a M + \log_a N$.

Proof

If M.N is the product and if $a^x = M$ and $a^y = N$, then $M.N = a^x . a^y$.

By the law of indices $a^x . a^y = a^{x+y}$. Therefore,

$a^{x+y} = M.N$

Then the logarithm of M.N to base a is equal to x + y. Mathematically, it will be

$\log_a M.N = x + y$ ..... (1)

Now, if we express $a^x = M$ and $a^y = N$, in terms of logarithms they will be $\log_a M = x$ and $\log_a N = y$. Substituting the values of x and y in 1, we have

$\log_a (M.N) = \log_a M + \log_a N$

Example 3

What is the value of $\log_3 33$?

We know that 33 can be expressed as the product of 3 and 11. That is, $\log_3 33 = \log_3 (3 \times 11)$. Applying the above rule this can be expressed as $\log_3 3 + \log_3 11$. Since $\log_3 3$ is 1, we rewrite it as $\log_3 33 = 1 + \log_3 11$.

Rule 4

The logarithm of a fraction to the base a will be equal to the ratio of the logarithms of the numerator to the base a and the logarithm of the denominator to base a. That is, $\log_a (M/N) = \log_a M - \log_a N$.

Proof

Let $a^x = M$ and $a^y = N$. Then $M/N = a^x/a^y$. By the law of indices, this will equal to a^{x-y}. The logarithm of M/N to base a will, therefore, be x – y. Mathematically this is expressed as

$$\log_a (M/N) = x - y \qquad .....(1)$$

If we express $a^x = M$ and $a^y = N$ in terms of logarithms, they will be $\log_a M = x$ and $\log_a N = y$. Substituting the values of x and y in (1), we have

$$\log_a (M/N) = \log_a M - \log_a N.$$

Example 4

What is the value of $\log_2 (32/4)$?

By applying the above rule, this can be written as $\log_2 32 - \log_2 4$. This can be further solved. But we look at it only after learning the next rule.

Rule 5

The logarithm of a number raised to any power, integral or fractional, is equal to product of that number and the logarithm of number which was raised to base a. That is, $\log_a (M^P) = p.\log_a M$.

Proof

If $M = a^x$, then $\log_a M = x$. Now suppose that M is raised to the power of n, that is M^n. Since $M = a^x$, $M^n = a^{nx}$. This is in accordance with the priniciple that if we perform any operation on an equation it should be performed on both the sides of the equation in order to keep the equation symbol valid.

$M^n = a^{nx}$, if expressed in terms of logarithms will be

$$\log_a(M^n) = nx \qquad .....(1)$$

On substituting $\log_a M = x$ in (1), we have

$$\log_a (M^n) = n \,.\, \log_a M$$

Similarly if n = 1/r, we have

$$\log_a (M^{1/r}) = (1/r) \,.\, \log_a M$$

Now we take up the example discussed under Rule 4 and look at how it is further simplified. Before we go on to the next step, let us express $\log_2 32$ and $\log_2 4$ as $\log_2 2^5$ and $\log_2 2^2$. By rule 5, these are expressed as $5.\log_2 2$ and $2.\log_2 2$. And since $\log_2 2$ is one, $5.\log_2 2$ and $2.\log_2 2$ reduce to 5.1 = 5 and 2.1 = 2. Therefore, $\log_2 32 - \log_2 4$ when simplified gives

$$\log_2(2^5) - \log_2(2^2)$$

$$= \quad 5.\log_2 2 - 2.\log_2 2$$

$$= \quad 5.1 - 2.1$$

$$= \quad 5 - 2 = 3.$$

We obtain the same value even by simplifying the term on the left hand side. We know that 32/4 = 8. That is, $\log_2 8.8$ can be expressed as 2^3. On application of rule 5, this will be $3.\log_2 2$. Again this gives us 3.1 = 3.

Generally, logarithms are expressed to base 10 and base e. While the logarithms expressed to base 10 are referred to as common logarithms, those expressed to base e are referred to as Napier or Natural logarithms. The value of e is approximately 2.718. In practise common logarithms are expressed as $\log_{10}300$ while natural logarithms are expressed as ln 40. We want to emphasize that generally the base is not stated and by looking at the manner it is expressed we ought to decide whether it is common or natural logarithm.

TRANSFORMING THE BASE OF LOGARITHMS

Suppose that we know the logarithms of all numbers which are expressed to base a and we are required to find the logarithms of all these numbers to base b. We proceed as follows. Let N be any one of the numbers of which we are required to find the logarithm to base b and the value itself be some x. That is, $\log_b N = x$ or $N = b^x$. But we already know the value of $\log_a N$. Also $\log_a N$ can be expressed as $\log_a(b^x)$ as $N = b^x$. By rule 5, $\log_a(b^x)$ can be expressed as $x.\log_a b$ or

$$x = \frac{1}{\log_a b} x \log_a N \text{ or}$$

$$\log_b N = \frac{1}{\log_a b} x \log_a N \quad \text{..... (1)}$$

Since the values of N and b are known, the values of $\log_a N$ and $\log_a b$ can be found from the tables. These values when substituted in equation (1) gives us the value of $\log_b N$.

In the above equation, what will happen if N = a.
Equation (1) will be

$$\log_b a = \frac{1}{\log_a b} x \log_a a = \frac{1}{\log_a b} \text{ (because } \log_a a = 1)$$

or

$$\log_b a \times \log_a b = 1.$$

Example 5

Find the values of the following:

a. $\log_3 81$

We know that $81 = 3^4$. Therefore, $\log_3 3^4 = 4. \log 3 = 4.1 = 4$

b. $\log_3 (9 \times 27 \times 81)$

$\log_3 (9 \times 27 \times 81) = \log_3 9 + \log_3 27 + \log_3 81$ ($\log_a M.N = \log_a M + \log_a N$)

$= \log_3(3^2) + \log_3(3^3) + \log_3(3^4)$

$= 2.\log_3 3 + 3.\log_3 3 + 4.\log_3 3$

$= 2 + 3 + 4 = 9$

c. $\log_3 (3)\frac{5}{4}$

In this example we apply rule 5.

$(5/4).\log_3 3 = (5/4).1 = (5/4)$

d. $\log_3 (243/81)$

$\log_3 243 - \log_3 81$ [$\log_a (M/N) = \log_a M - \log_a N$]

$\log_3 3^5 - \log_3 3^4$

$5.\log_3 3 - 4.\log_3 3$ ($\log_M (M^p) = p.\log_M M = p$)

$=5 - 4=1$

SUMMARY

- The power to which a base (say 10) should be raised to produce a given number is called Logarithm.
- The study throws light on the various rates pertaining to Logarithms change of base of Logarithms etc.

Annexure

A Note on the Sigma (Σ) Notation

In statistics many important formulae like the mean and standard deviation involve the totaling of numbers. In fact there is a symbolic way to represent such totaling.

Suppose there are four numbers to be totaled. They may be represented as x_1, x_2, x_3 and x_4. The total of the four numbers $x_1 + x_2 + x_3 + x_4$ may also be written as

$$\sum_{i=1}^{4} x_i$$

The symbol sigma ($\sum$) indicates a summation or totaling. The x_i's are the numbers to be totaled. The "i=1" below $\sum$ and the "4" above $\sum$ indicate that the four numbers to be totaled are x_1 [i.e. x_i when i = 1], x_2 [i.e. x_i when i = 2], x_3 [i.e. x_i when i = 3] and x_4 [i.e. x_i when i = 4]. In short i takes values from 1 to 4

So $\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$

In general,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$$

Very often $\sum_{i=1}^{n} x_i$ is written as $\sum x$ with the understanding that all the values taken by x have to be totaled.

Hence, if x is a variable taking the values –3, 0, 7, 8 then $\sum x = -3 + 0 + 7 + 8 = 12$.

Some Properties of Σ

1. $\sum_{i=1}^{n} k = nk$ if k is a constant

   For example, $\sum_{i=1}^{3} 6 = 6 + 6 + 6 = 18 = 3 \times 6$

   [In this case, n = 3 and k = 6].

2. $\sum_{i=1}^{n} kx_i = k \sum_{i=1}^{n} x_i$

   For example, if x takes the values 2, 4 and 5 and k = 4 we have

   $\sum_{i=1}^{n} kx_i = \sum_{i=1}^{3} 4x_i = (4 \times 2) + (4 \times 4) + (4 \times 5)$

   $= 44 = 4 \times 11 = 4 \times (2 + 4 + 5) = 4 \sum_{i=1}^{3} x_i = k \sum_{i=1}^{n} x_i$

3. $\sum (x + y) = \sum x + \sum y$ where x and y are two variables.

   For example, if x and y take the following values, we have

				Total
	3	6	15	24
y	8	1	7	16
x + y	11	7	22	40

   $\sum (x + y)) = 11 + 7 + 22 = 40 = (3 + 6 + 15) + (8 + 1 + 7) = \sum x + \sum y$

4. The arithmetic mean of a variable x is represented by $\frac{\sum x}{N}$

   Where, N is the number of observations or data points.

5. The standard deviation of a variable x is represented by

$$\sigma = \frac{\sqrt{\sum (x - \mu_x)^2}}{N}$$

Where, μ_X is the mean of the variable x and N is the number of data points in the population.

We can simplify this formula as below:

$$\sigma = \sqrt{\frac{\sum (x^2 - 2x\,\mu_x + \mu_x^2)}{N}}$$

[because $(x - \mu_X)^2 = x^2 - 2x\,\mu_X + \mu_X^2$]

$$= \sqrt{\frac{\sum x^2 - 2\mu_x \sum x + N\mu x^2}{N}}$$

[because $2\mu_X$ is a constant

$\sum 2x\,\mu_x = 2\mu_x \sum x$ and $\sum \mu_X^2 = N\mu_X^2$]

$$= \sqrt{\frac{\sum x^2 - 2\mu_x N\,\mu_x + N\mu_x^2}{N}}$$

$$= \sqrt{\frac{\sum x^2 - N\mu_x^2}{N}}$$

$$= \sqrt{\frac{\sum x^2}{N} - \mu_x^2} = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$$ [because $\sum x = N\,\mu_x$]

Similarly in the case of the standard deviation S of sample we have

$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ where, x is the variable.

x is the sample mean and n is the number of sample points (data points in the sample)

The above formula can be reduced to

$$S = \sqrt{\frac{\sum x^2}{n-1} - \frac{n\bar{x}^2}{n-1}}.$$

Chapter II

Introduction to Statistics

After reading this chapter, you will be conversant with:

- Origin and Growth of Statistics
- Applications of Statistics
- Collection of Data
- Significance of Computers in Statistics

Introduction

The word statistics is not new for human society. It has been used right from the existence of life on earth. Though its use was much limited in the ancient days it was regarded as the 'Science of Statecraft' and was the by-product of the administrative activity of the State. It has been the traditional function of the governments to keep records of population, births, deaths, taxes, crop yields, and many other types of activities. The demand for statistics has been growing considerably since the past century and its significance is recognized in recent years with the enormous development taking place in the field of business and commerce, governmental activities, etc. It helps in the formulation of policies and procedures.

The expansion of business activities has made the operations of business more complex. These complexities can be resolved by gathering factual information, which is done through statistics. The statistical procedures help in resolving business complexities.

Different writers have defined statistics differently from time to time. The most exhaustive definition given by Prof. Horace Secrist is like –

"By Statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other".

i. **Aggregate of Facts:** Single or isolated figures cannot be termed as Statistics unless they are a part of aggregate of facts relating to any particular field of enquiry. However, aggregate of the figures of births, deaths, sales, purchase, production, profits etc., over different times, places, etc., will constitute Statistics.

ii. **Affected to a Marked Extent by Multiplicity of Causes:** Generally, the facts and figures are affected to a considerable extent by a number of forces operating together. For example, the prices of a particular commodity are affected by a number of factors, such as supply, demand, imports, exports, money in circulation, competitive product in the market and so on. Similarly the yield of a particular crop depends upon multiplicity of factors like, quality of seed, fertility of soil, method of cultivation, irrigation facilities, weather condition, fertilizer used and so on. It is very difficult to study separately the erect of each of these forces. In physical sciences, it is possible to isolate the effects of various forces on a particular event, but it is very difficult in case of social sciences, as the factors cannot be measured quantitatively. However, statistical techniques like, Multiple correlation and partial correlation have been devised to study the joint effects of number of factors on a single item.

iii. **Numerically Expressed:** Numerical facts only constitute Statistics. Qualitative statements like, 'India is a poor country', 'the production of wheat is increasing' do not constitute statistics. Further, the qualitative characteristics which cannot be measured quantitatively such as intelligence, beauty, honesty, etc., cannot be termed as Statistics unless they are numerically expressed by assigning particular scores as quantitative standards.

iv. **Enumerated or Estimated According to Reasonable Standard of Accuracy:** The numerical data pertaining to any field of enquiry can be obtained by complete enumeration. Where complete enumeration is not possible due to large population, high cost of enumeration per unit and limited time and resources, then the data are estimated by using sampling and estimation techniques. However, the estimated value will not be as precise and accurate as the actual values. The degree of accuracy of the estimated values largely depends on the nature and purpose of the enquiry. For example, in measuring heights of persons accuracy will be aimed in terms of

fraction of an inch whereas in measuring the distance between two places, say, Delhi and Chennai, even fraction of a kilometer can be ignored. However, it is important that the reasonable standards of accuracy must be maintained for drawing meaningful conclusions.

v. **Collected in a Systematic Manner:** A suitable plan of data collection should be prepared before collecting statistics and the work is carried out in a systematic manner. An attempt should be made to reduce the personal bias to the minimum. The data collected in a haphazard way will not confirm to the reasonable standard of accuracy and the conclusions based on them might lead to fallacious conclusions.

vi. **Collected for a Pre-determined Purpose:** The purpose of collecting data must be decided in advance. The purpose should be specific and well-defined. An attempt should not be made to collect too many data some of which are never examined or analysed and also it should be ensured that no essential data are omitted. For example, if the purpose of enquiry is to measure the cost of living index for low-income group people, we should select only those commodities or items, which are consumed or utilized by persons belonging to this group.

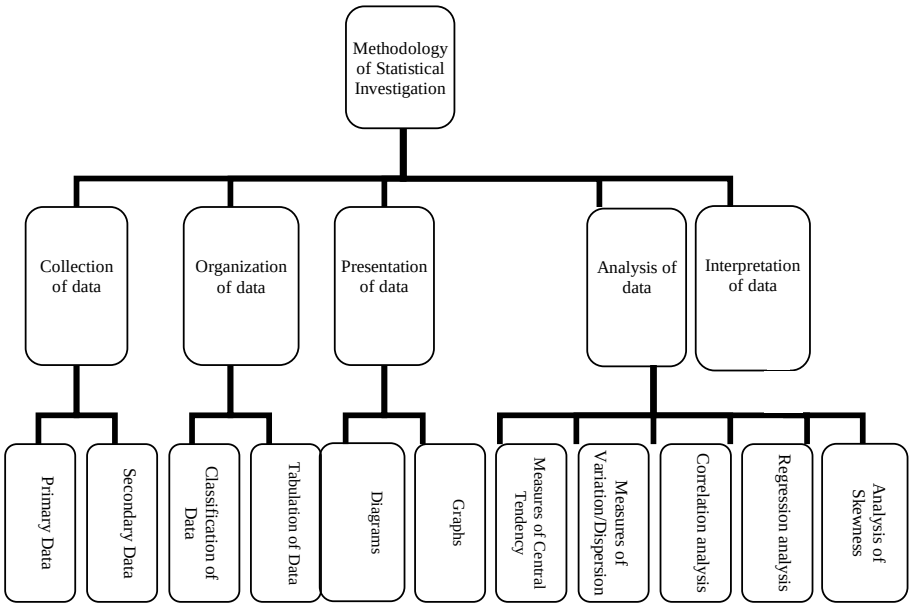
vii. **Comparable:** For statistical analysis, the data should be comparable. Statistical data are often compared period-wise or region-wise. For example, the population of India at a particular point of time may be compared with that of earlier years or with the population of other countries. However, it would be meaningless to compare the data relating to the size of shoe of an individual and his I.Q. A valid comparison can be made only if the data are homogeneous, i.e., relate to the same phenomenon or subject.

Methodology of Statistical Investigation

The above definitions are much narrower and incomplete. They exhibit only the features of statistics but statistics also includes different stages. Hence, the following definition given by Croxton and Cowden finds to be apt in this context.

According to Croxton and Cowden statistics may be defined as the "collection, presentation, analysis and interpretation of numerical data".

Figure 1



These stages can be discussed as below:

- Collection of data is the first step in statistics. The reliable and appropriate data forms the life blood for statistical inferences. Again, the data is classified as primary data and secondary data. Varied sources of collection viz., sources of primary data and secondary data form the subject matter here. This can be discussed in detail in later pages of the chapter.
- Organization of data is the second step. The data so collected will be placed haphazardly in the first step. Hence its organization plays an important role in this context. Classification and tabulation are the important tools used in this process. Sometimes data collection involves the collection of large mass of figures. These should be organized to get the clear picture about the data collected. This aspect has been discussed in subsequent chapters.
- The next step in the statistical analysis is the presentation of data. This can be done through diagrams and graphs. This aspect has been discussed in subsequent chapters.
- Analysis of the data forms the subject matter of the statistics. The utility of various statistical tools like measures of central tendency, variation/dispersion, correlation and regression analysis, skewness etc., comes at this stage. These are discussed in later chapters.
- Interpretation of the data so analyzed is the final step in statistical analysis. Without interpretation the analysis goes waste. Hence it should be done thoroughly.

Such data can be widely used for business, education, research, etc., purposes.

ORIGIN AND GROWTH OF STATISTICS

The word 'statistics' comes from the Italian word 'statista' which means Statement or a German word 'statistik' which means a Political State. It was first used by Prof. Gottfried Achenwall (1719-1772), a professor in Marlborough 1749 which refers to the subject matter as a whole. The word 'statistics' appeared for the first time in the famous book, 'Elements of Universal Erudition' by Baran J.F. Von Bielfeld, translated by W. Hooper M.D.

Statistics is now regarded as one of the most important tools for taking decisions in the midst of uncertainty. In fact, there is hardly any branch of science today that does not make use of statistics. In the present century, considerable development has taken place in the fields of business and commerce, Governmental activities and Science. In case of business, not only the magnitude of business has increased considerably but also growing size of business has made its problems more complex. Most of these problems are resolved in the light of factual information for which statistics is essential. Now-a-days, the activities of the government are also increased unlike the time when the primary duty of the Governments was to maintain law and order. The demand for statistics has increased due to these enlarged activities. For the tremendous development of Science, as a tool of research, the importance of statistics is innumerable. With the development of electronic machines, the cost of analyzing data also gone down. Further, with the development of statistical theory, the cost of collecting and processing data has also gone down. This has led to the increasing use of statistics in solving various problems.

APPLICATIONS OF STATISTICS

Before discussing the applications of statistics it is advisable to discuss the various functions of statistics as they provide clarity regarding the applications. These functions are discussed as under:

Functions of Statistics

Statistics is primarily concerned with the analysis and interpretation of collected and organized data. In this process it has certain functions to be performed. These functions are beautifully summed up Robert W. Burgess in the following way:

"The fundamental gospel of statistics is to push back the domain of ignorance, rule of thumb, arbitrary, or premature decision, traditions and dogmatism and to increase the domain in which decisions are made and principles are formulated on the basis of analyzed quantitative facts".

This definition significantly expresses that the statistics has certain functions/principles to be performed. They are discussed below:

i. One of the important functions of statistics is to present general statements in a precise and definite form. It establishes clarity in its statements through this function.

ii. It simplifies the data by condensing the unnecessary information. Thus, only material information is presented to the user of the data.

iii. It facilitates comparison of data/figures with their respective counterparts e.g., sales and expenditure, wages and consumption etc., thus, enabling to draw inferences for a specific purpose.

iv. Formulation of hypothesis and testing the same is another function of statistics. For example, to check that the credit contraction is effective in checking the increase in prices hypothesis testing is applied. Thus, it helps in the development of new theories.

v. Statistical methods aid in forecasting or predicting future trends. Thus, business organizations use statistical tools in preparing and implementing future plans and policies, budgets etc.

vi. It aids its various users in the preparation of suitable policies by providing basis material to them. For example, by providing data about population statistics help in the distribution of population by age, sex and other socio – economic characteristics, which in turn help in determining future needs such as food, clothing, etc.

Applications of Statistics

The data obtained by using the above discussed statistical methodology applied or used by many. These applications/advantages can be studied as under:

STATISTICS AND STATE

Today statistical data relating to prices, production, consumption, income and expenditure, investments and profits etc., and statistical tools such as index numbers, time series analysis, demand analysis, forecasting etc., are extensively used by the governments in formulating economic policies. The use of statistical data and statistical techniques is so wide in government functioning that today almost all the ministries and departments in the government have a separate statistical unit and in most of the countries, State (government) is the single unit which is the biggest collector and user of statistical data.

STATISTICS AND BUSINESS

Before Industrial Revolution, the business activities were very much limited. But after the revolution, the developments in business activities taken unprecedented dimension both in size and competition in the market. With the increasing competition, the problems of the business enterprises are becoming complex and they are using more and more statistics in decision-making. Management has

become a specialized job and a manager is called upon to plan, organize, supervise and control the operations of the business house. Most of the production these days is in anticipation of demand and, therefore unless a very careful study of the market is made, the firm may not be able to make profits. Thus a businessman who has to deal in an atmosphere of uncertainty, has to deal systematically with the uncertainty itself by careful evaluation and application of statistical methods, concerning the business activities. Business indeed runs on estimates and probabilities. The higher the degree of accuracy of a businessman's estimates, the greater is the success attending on his business. It has become increasingly evident that statistics and statistical methods have provided the businessman with one of his most valuable tools for decision-making. Time Series Analysis, Index Numbers, Forecasting Techniques and Demand Analysis are some of the very important powerful statistical tools, which are used immensely in the analysis of economic data and also for economic planning.

STATISTICS AND ECONOMICS

Prof. Alfred Marshal, the renowned economist, observed the significance of statistics in economics in 1890 that "Statistics are the straw out of which I, like every other economist, have to make bricks". Statistical data and advanced techniques of statistical analysis have proved immensely useful in the solution of a variety of economic problems such as production, consumption, distribution of income and wealth, wages, prices, profits, savings, expenditure, investment, unemployment, poverty etc. In recent years, econometrics, which comprises the application of statistical methods to the theoretical economic methods, is widely used in economic research. The questions like, calculation of national income and its distribution cannot be answered without statistics. In reducing disparities in the distribution of income and wealth statistics are of immense help. Similarly, in solving problems of rising prices, growing population, unemployment, poverty, etc., one has to rely heavily on statistics. Most of the economic policies would be a leap in the dark in the absence of appropriate statistical information.

STATISTICS AND RESEARCH

The usage of statistics is inevitable in research work. Today, most of the research work is done by using statistical tools. The usage of statistics can be done in the fields of agriculture, medicine, population analysis, education, etc. This can be clearer from the following example:

This example is a case of agriculture; a farmer can better utilize the fertilizers only when he gets acquainted with the benefits conferred by them. The green revolution is the outcome such through research analysis. Similar is the case with other fields.

STATISTICS AND OTHER USES

Besides the above, statistics are useful to bankers, brokers, insurance companies, social workers, labor unions, trade associations, chambers of commerce and to the politicians. For example, the success of an insurance company largely depends on the accuracy of the statistical data, which is used to study the life expectation for fixing the premium rates. Similarly, prior to election, different political parties use sampling methods to know the winning chance of the candidates, and accordingly they decide the campaign effort for its success. The banks have to make a very careful study of the cash requirements otherwise they may find themselves short of cash and their existence is at stake.

Limitations of Statistics

i. **Statistics does not Study Individuals:** According to the definition a single or isolated figure cannot be regarded as statistics unless it is a part of the aggregate of facts relating to any particular field of enquiry. Thus, statistical methods do not give any recognition to an object or a person or an event in isolation. For example, the wages of workers of a factory can be statistics, but the wage earned by an individual worker at any one time taken by it is not a statistical datum. But, Similarly the marks obtained by one student of a class or his height is not the subject matter of the study of statistics but the average marks or the average height has statistical relevance.

ii. **Statistics deals with only Quantitative Characteristics:** Statistics, as a science, deals with a set of numerical data that can be applied to the study of only those phenomena, which can be measured quantitatively. As such statistics cannot be used directly for the study of qualitative characteristics like, honesty, efficiency, intelligence, blindness and deafness etc., however, it may be possible to analyse such problems statistically by expressing them numerically after assigning particular scores or quantitative standards.

iii. **Statistical Laws are not Exact:** Statistical laws are true only on an average. The conclusions obtained statistically are not universally true; they are true only under certain conditions. This is because, statistics as a science, is more approximate and less exact as compared to natural sciences.

iv. **Statistics is Liable to be Misused:** The most significant limitation of statistics is that it is liable to be misused. Statistics is merely a tool which deals with figures which are innocent in themselves and do not bear on their face the label of their quality and can be easily distorted, manipulated, or moulded by politicians, dishonest or unskilled workers, unscrupulous people for personal selfish motives. Statistics when used rightly may prove extremely useful but if misused by inexperienced, unskilled and dishonest statisticians might lead to very fallacious conclusions and even prove to be disastrous.

Distrust of Statistics

Distrust of statistics means lack of confidence in statistical statements and statistical methods. People often comment, “Statistics can prove anything”. The main reasons of such comments about statistics can be that statistical figures are convincing and they can be manipulated in such a manner as to establish foregone conclusions. Statistics is a tool or a method of approach. If these tools are used properly they will help in taking wise decisions and if misused, they can do more harm than good. But the fault does not lie with the science of statistics. For example, medicines are meant for curing people, but if a wrong medicine is taken or an excessive dose of medicine is taken a person may die. Thus, if statistical facts are misused by some people it would be wrong to blame the science. It is the people who are to be blamed. In fact, statistics is like clay, of which one can make a God or Devil as he pleases.

COLLECTION OF DATA

It is, as said above is the primary stage in statistical investigation. For any statistical enquiry, whether it is in business, economics or social sciences, the basic task is to collect facts and figures relating to particular phenomenon under study. The person who conducts the statistical enquiry is known as investigator. The process of counting or enumeration or measurement together with the systematic recording of results is called the collection of statistical data. Data may be obtained either from the primary source or the secondary source. When the data collected originally by the investigator for the given enquiry, it is called primary data. The data, which are not originally collected but rather obtained from published or unpublished sources, is known as secondary data.

The difference between primary and secondary data is only of degree. The data, which are primary in the hands of one, become secondary in the hands of another. Data are primary for the individual agency or institution collecting them whereas for the rest of the world they are secondary.

Using of secondary data have the following advantages:

i. If secondary data is available they are much quicker to obtain than primary data.

ii. In some cases, it would be impossible to collect primary data. For example, Census data cannot be collected by an individual or research organization, but can be obtained from Government publications.

However, the major problems in using secondary data are the difficulty of finding data, which exactly fit the need of the present project and finding the sufficient data on a particular area.

It is the investigator who will decide whether he will use primary data or secondary data in an investigation. The choice between the two depends on the following considerations:

a. Nature and scope of the enquiry.

b. Availability of financial resources,

c. Availability of time,

d. Degree of accuracy desired, and

e. The collecting agency, i.e., whether an individual, an institution or a government body.

Methods of Collecting Primary Data

DIRECT PERSONAL INTERVIEW

In this method, the investigator collects data by visiting the field personally and face-to-face contact with the persons for making enquiries and soliciting the information from the informants or respondents. The information thus obtained is first-hand or original in character. This method is suitable, where intensive study of a limited field is desired.

Merits

i. The first-hand information obtained by this method is likely to be more accurate since the investigator can remove the doubts if any in the minds of the respondents and can extract the correct information.

ii. When approached personally, most people will supply information, thus resulting more encouraging response for the study.

iii. Through personal interview, some supplementary information can be obtained which often proves very useful while interpreting results.

iv. When the investigator will feel that the information is foul, he can check it by some intelligent cross-questions.

v. The investigator can get proper information, the language of communication can be adjusted to the status and educational level of the person interviewed, thus avoiding inconvenience and misinterpretation on the part of the informant.

vi. Sensitive questions can be carefully asked by twisting the question to extract proper information.

Demerits

i. It is one of the costly methods of data collection as the number of persons to be interviewed is large and they are spread over a wide area.

ii. It is a time consuming process, which may restrict the purpose of the study.

iii. The success of the investigation largely depends upon the the intelligence, skill, tact, insight, diplomacy and courage of the investigator. An investigator without these qualities and a poorly trained one may spoil the entire work.

iv. The personal prejudice, bias and whims of the investigator in certain cases may affect the findings of the enquiry.

INDIRECT ORAL INTERVIEWS

Under this method, the investigator contacts third parties called witnesses capable of supplying the necessary information. This method is generally used in those cases where the information to be obtained is of a complex nature and the informants are not inclined to respond if approached directly. For example, when

somebody wants to solicit information on certain social evils, and while collecting information from addicted people regarding drinking, gambling, or smoking etc., the people may be reluctant to supply information or may give wrong information about their own hobbies. In this case, it would be necessary to get the information from those dealing in drugs, liquor or other people who may be their neighbors, relatives or personal friends. This method is very popular in practice. The police obtain clues about the theft or murders by interrogating third parties who are supposed to have knowledge about the case under investigation. The correctness of information obtained depends upon the factors such as:

a. The type of persons whose evidence is being recorded, if the people do not know the full facts of the problem or if they are prejudiced, it will not be possible to arrive at correct conclusions.

b. The ability of the interviewers to draw out the information from witnesses by means of appropriate questions and cross-examination.

c. The honesty of interviewers who are collecting the information.

Merits

i. The enumerator can exercise their intelligence, skill, tact etc., to extract correct and relevant information by cross-examination as the enumerators contact the informants personally.

ii. This method is less expensive and requires less time as compared to direct personal interview method for conducting the enquiry.

iii. In order to formulate and conduct the enquiry more effectively and efficiently, the expert views and suggestions of the specialists on the given problem can be obtained.

INFORMATION FROM CORRESPONDENTS

In this method, the investigator appoints local agents or correspondents in different places to collect information. These correspondents or agencies in different regions collect the information according to their own ways, fashions, likings, and decisions and then submit their reports periodically to the central or head office where the data are processed for final analysis. This method of data collection is usually employed by Newspaper agencies who require information in different fields like sports, riots, strikes, accidents, economic trend, business stock and share market, politics and so on. Various departments of government in such cases also adopt this technique where regular information is to be collected from a wide area. This method is particularly suitable in case of crop estimates. This method generally pertains where the information is to be obtained at regular intervals from a wide area.

Merits

i. This method is cheap and appropriate for extensive investigation.

ii. The required information can be obtained expeditiously since only rough estimates are required.

Limitations

The data collected by this method is not very reliable, as the different correspondents collect the information in their own fashion and style; the results are bound to be biased due to personal prejudices and whims of the correspondents in different fields of the enquiry.

MAILED QUESTIONNAIRE METHOD

Under this method, a questionnaire is prepared and is mailed to various informants by post. The questionnaire contains list of questions relating to the field of enquiry and providing space for the answers to be filled by the respondents. A polite covering note, explaining in detail the aims and objectives of collecting the information and also the operational definitions of various terms and concepts used

in the questionnaire is attached. Respondents are also requested to extend their full cooperation by furnishing the correct replies and returning the questionnaire duly filled in time. The respondents are also taken into confidence by ensuring them that the information supplied by them will be kept strictly confidential and secret. To ensure quick and better response, usually the investigator sends a self-addressed stamped envelope. The questions asked in the questionnaire should be clear, brief, corroborative, non-offending, and courteous in tone, unambiguous and to the point so that not much scope of guessing is left on the part of the respondent. This method is usually used by the research workers, private individuals, non-official agencies and sometimes by Central or State Government.

Merits

i. This method of collection of data is considered as the most economical method in terms of time, money and manpower and particularly, when the field of investigation is very vast and the informants are spread over a wide geographical area.

ii. Errors due to the personal biases of the investigators or enumerators are completely eliminated as the information is supplied directly by the person concerned in his own handwriting. The information so obtained is original and much more authentic.

Limitations

i. This method can be adopted only where the informants are literate, so that they can understand written questions and send the answers in writing.

ii. Sometimes, people might suppress correct information and furnish wrong replies. It is very difficult on the part of investigator to verify the accuracy and reliability of the information received.

iii. Informants may not be willing to give written information in their own handwriting on certain personal questions like, income, property, personal habits and so on.

iv. It involves some uncertainty about the response; Co-operation on the part of the informants may be difficult to presume.

Drafting the Questionnaire

Draft questionnaire is an important aspect of this method. It is the focal aspect of this method. Questionnaire is the only media of communication between the investigator and the respondents. Hence, it should be designed or drafted with utmost care and caution. Drafting of a good questionnaire is a highly specialized job and requires great care, skill, wisdom, efficiency and experience. It is difficult to lay down any hard and fast rules to be followed for designing a questionnaire. However, the following general principles may be helpful in framing a questionnaire:

i. **The Size of the Questionnaire should be Small:** The number of questions should be kept to the minimum, keeping in view the nature, objectives, and scope of the enquiry. If the number of questions in a questionnaire is more than 15 or 20 it should be preferably be divided into two or more parts. There is an inverse relationship between the length of a questionnaire and the rate of response to the survey. That is; the longer the questionnaire, the lower will be the rate of response; the shorter the questionnaire, the higher will be the rate of response.

ii. **The Questions should be Short and Simple to Understand:** The questions should be short and simple and technical terms should be avoided.

iii. **The Questions should be Arranged Logically:** The questions must be arranged in a logical order thus enabling natural and spontaneous reply to each question. For example, it would be illogical to ask a man his income before asking him whether he is employed or not. Thus, the sequence of the question should be considered carefully in terms of the purpose of the study.

iv. **Personal Questions should be Avoided:** Questions of sensitive and personal nature should be avoided. For example; questions about income, savings, sales-tax paid etc., on which the respondents may be reluctant or unwilling to furnish information. Similarly the questions, which might hurt the sentiments, pride or prestige of an individual should not be asked as far as possible.

v. **The use of Vogue and 'Multiple Meaning' words should be Avoided:** The vogue words like good, bad, efficient, sufficient, prosperity, rarely, frequently, reasonable, poor, rich, etc., should not be used since these may be interpreted differently by different persons and may give unreliable and misleading information. Similarly, the use of words with multiple meanings like, price, assets, capital, income, household, democracy, socialism etc., should not be used unless a clarification to these terms is given in the questionnaire.

vi. **Questions should be Capable of Objective Answer:** The questions asked in the questionnaire should be short, simple alternatives, and preferably multiple-choice questions.

vii. **Questions Requiring Calculation should be Avoided:** The questions requiring calculations should be avoided. For example, questions necessitating calculation of ratios and percentages, etc., should not be asked which may take much time and the informant may not send back the questionnaire.

SCHEDULES SENT THROUGH ENUMERATORS

In this method, schedules are being send through the enumerators or interviewers to get replies to the questions contained in a schedule and fill them in their own handwriting in the questionnaire form.

Merits

i. This method can be adopted in the cases, where the informants are illiterate. Because, the enumerators can explain in detail the objectives and aims of the enquiry to the informants and impress upon them the need and utility of furnishing the correct information.

ii. There is very little non-response as the enumerators go personally to obtain the information.

iii. The information received by using this technique is more reliable as the accuracy of statements can be checked by supplementary questions wherever necessary.

Limitations

i. It is an expensive method and time consuming as compared with 'mailed questionnaire method'.

ii. The success of the method depends largely upon the efficiency and skill of the enumerators.

iii. For making the interview skilled, it requires experience and training of the enumerator. They should be well versed in the local language, customs and traditions. Without good interviewing most of the information collected is of doubtful value.

iv. Due to inherent variation in the individual personalities of the enumerators there is bound to be variation, though not so obvious, in the information recorded by different enumerators.

v. When the schedule is framed haphazardly and incompetently, the enumerator will find it very difficult to get the complete and correct desired information from the respondents.

Sources of Secondary Data

- Published sources, and
- Unpublished sources.

PUBLISHED SOURCES

There are a number of organizations and agencies at national and international level, which collect statistical data relating to different matters and publish their findings in statistical reports on a regular basis. These publications serve as a powerful source of secondary data. The various sources of published data are:

1. Official Publications of
    a. International bodies such as the 'World Bank', 'International Labor Organization', 'World Health Organization', 'International Monetary Fund' etc.,
    b. Central and State Governments such as Economic Survey, Ministry of Finance, Government of India,
    c. Reports of the Committees and Commissions appointed by the Government, such as Fifth Pay Commission Report, Kothari Commission Report on Educational Reform, Land Reform Committee Report, etc.
2. Publications of Semi-Government Statistical Organizations like,
    a. Economic department of RBI.
    b. The Institute of Foreign Trade.
    c. Gokhale Institute of Politics and Economics, and
    d. Publications of various local bodies such as Municipal Corporation and District Boards.
3. Publications of autonomous and private Research Institutions.
    a. Publications brought out by various autonomous research organizations like Indian Statistical Institute, Indian Council of Agricultural Research, National Council of Educational Research and Training (N.C.E.R.T.), National Council of Applied Economic Research (N.C.A.E.R.) etc.
    b. Publications of private commercial, trade and professional bodies like Federation of Indian Chamber of Commerce and Industry, the Institute of Chartered Accountants, Trade Unions, Stock Exchanges, etc.
    c. Financial and Economic Journals such as The Indian Journal of Economics, Commerce, Reserve Bank of India Bulletin, etc.

UNPUBLISHED SOURCES

There are various sources of unpublished statistical data such as records maintained by private firms or business enterprises, the various departments and offices of the Central and State Governments, the researches carried out by the individual research scholars in the universities or research institutes.

Editing Primary and Secondary Data

Editing primary or secondary data is the next step in data collection. The main objective of the step is to detect any errors in the data so collected. It is one of the crucial steps taken and requires considerable care and diligence. Any little negligence makes the whole act of editing useless.

While editing the primary data the following points need to be considered:

- The completeness of data, i.e., the editor should see that all the schedules in the data are complete and no section is left vague.

- The editor should perceive that the content in the data is consistent i.e., it involves no contradictory statements.
- The reliability of conclusions depends basically on the correctness or accuracy of information/data. Any inaccurate or wrong statement leaves the data invalid. Hence accuracy of data is one of the important considerations.
- The editor must also perceive that the entire information is interpreted in a uniform manner. The editor should check that the data obtained from various sources is homogeneous and uniform.

Precautions in the Use of Secondary Data

While using the secondary data, the investigator must be satisfied regarding the reliability, accuracy, adequacy and suitability of the data to the given problem under investigation.

i. **Reliability of Data:** It is very difficult to know whether the secondary data are reliable or not. The data collector should satisfy with the following points:
   a. The reliability, integrity and experience of the collecting organization.
   b. The methods used for the collection and analysis of the data.
   c. If the enumeration was based on a sample; whether the sample is adequate and representative of universe?
   d. The editing, tabulating and analyzing is done carefully and conscientiously.
   e. The degree of accuracy desired by the compiler was achieved or not.

ii. **Suitability of Data:** Even though the data is reliable, it should not be used without ensuring the suitability of the data for the purpose of enquiry. The suitability of data for the purpose of enquiry is important to observe and compare the objective, nature and scope of the investigation. Sometimes, the secondary data do not satisfy immediate needs because they have been compiled form other purpose. The value of secondary data is frequently affected by variation in the units of measurement such as; consumer income and variation in the date/period to which the data is related.

iii. **Adequacy of Data:** Adequacy of data is to be judged in the light of the requirements of the survey and the geographical area covered by the available data. Because, when the coverage given in the original enquiry is too narrow or too wide than what is desired in the current enquiry or on the other hand when the original data refers to an area or a period which is much larger or smaller than the required one. For example; if the object of a study is to know the wage rates of the workers in sugar industry in India, it would not serve the purpose if the data available covers only the State of U.P. Another important factor to decide about the adequacy of the available data for the given investigation is the time period for which the data are given. For example, for studying trend of prices we may use data for the last 8-10 years but the data may be available for the last 2-3 years only, which would not serve the purpose.

SIGNIFICANCE OF COMPUTERS IN STATISTICS

Computers are the new technological revolution and has its share of importance in every field; be it a business, economics, politics physical or social sciences, mathematics or statistics. It solves major problems in mere seconds viz., complex calculations, data processing etc.

As discussed above, statistics is not new to computer technology and is closely related to the evolution of electronic computing machinery. Statistics involves the conversion of data into information that aids in decision-making. Statistical

theories involve complexities and the calculations have been made more complicated. As such, the use of computers has become inevitable. As statisticians develop newer ways of describing and utilizing data for decisions, computer scientists, parallel to these developments, come out with newer and efficient means to perform these operations.

Though statistics uses computers to a considerable extent, the study of statistics need not involve the study of computers. Computers are only tools that aid in overcoming the statistical problems, if any. Computers and statistics are different disciplines and are poles apart in their nature and functions.

Inspite of the differences it can be said that the computer has made it possible to undertake statistical studies that involve the compilation and analysis of large masses of data and variables. The computations can be performed easily and quickly with the help of computer particularly when standard or packaged programs are available.

SUMAMRY

- Statistics is the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated and estimated according to reasonable standard of accuracy's, collected in a systematic manner for a predetermined purpose and placed in relation to each other.
- Statistics has several functions to be performed and is applied.
- Statistics is used not only in business and economics but for natural science and physical science also. It is a tool or method of approach for analysis.
- Statistics can be misused, if these tools are not used properly and can do more harm than good.
- Statistics uses only quantitative aspects of the data. Again it is not the study of individuals rather the study of aggregate of any field of enquiry.
- For the purpose of statistical analysis, data can be collected mainly in two ways such as primary source and secondary source.
    i. Primary source includes direct personal interview, indirect oral interview, mailed questionnaire method and scheduled sent through enumerators etc.
    ii. The secondary sources of data are published sources such as official publications of International bodies, publications of semi-government organizations, publications of autonomous and private research institutions, and unpublished sources such as records maintained by private firms on business enterprises, the researches carried out by the individual scholars in the universities or research institutes etc.
- Computers have significant role to play in statistics as they help in solving several statistical problems.

Chapter III

Sampling

After reading this chapter, you will be conversant with:

- Census and Sample Method
- Theoretical Basis of Sampling
- Methods of Sampling
- Size of Sample
- Merits and Limitations of Sampling
- Sampling and Non-sampling Errors

Introduction

In the previous chapter, we learnt that if the secondary data is not available for studying a problem, then a decision to collect primary data may be taken by using any one of the several methods discussed earlier. We can use the census and sample method for obtaining the required information. Population refers to the aggregate or totality of all members or items in the field of enquiry. In other words, complete enumeration of all items constitutes Universe or Population. A sample is the selected part that is used for determining the characteristics of the population. In other words, sample is the selection of respondents who are representative of total population. The process of selecting the respondents is called as sampling technique. The survey conducted is known as sample survey. Population size refers to total number of members of the population. The number included in the sample is known as sample size.

CENSUS AND SAMPLE METHOD

Census Method

Population is made up of several units. When every unit or item is approached for collecting the data, it is called as Census Method. In other words, when the data is collected for each and every unit of the population or the universe, is called Census Method. It is also known as Complete Enumeration Survey Method. In statistics, "population is the aggregate of object, animate or inanimate under study in any statistical investigation". The population may be finite or infinite population. A finite population is one, which has a finite number of items such as number of students in a college, number of industries in a country, etc. When the number of object or items are not finite it is termed as infinite population. For example, if we want to calculate the average wages of workers working in a cement factory, then we have to collect the wages of each and every worker working in a cement factory and divide this total by the number of workers working in cement factory in order to obtain the figure of average wage. The census method has following merits:

MERITS

i. Data is collected from each and every unit of the population or universe.

ii. The output or result is considered to be more representative, accurate and reliable.

iii. It is most suitable when information is to be collected on rare events.

iv. It provides the base for other surveys.

LIMITATIONS

However, this method is not widely used because of the following limitations inherent in it:

i. It is time consuming, requires more money and efforts.

ii. Sometimes the cost of conducting the census method is so high that the idea of collecting the information through this method is eliminated. It is practically true in case of underdeveloped countries.

iii. This method cannot be applied where the size of the population is large or where the evaluation process destroys the population unit.

iv. It is regarded needless if a sample yields equally reliable results.

Due to these limitations, another method known as sampling method has been developed.

Sampling Method

Sampling is the process of learning about the population on the basis of sample drawn from it. In the other words, it is the finite subset of the population, selected with the objective of investigating its properties. So in sampling, we study only a part of population instead of each and every unit of it. The theory of sampling is

quite old. Since immemorial times, a cook examines two or three grains of boiling rice to know whether it is ready or not. Another example of sampling is that by putting few questions to the students the teacher finds out whether the entire class has followed the lesson or not. Thus, sampling helps in knowing the characteristics of the population by examining a part of it.

OBJECTIVES

The main objectives of sampling are:

i. For obtaining the optimum results about the properties of the population with the available resources at its disposal.

ii. Obtaining the best estimates of the population parameter.

iii. For obtaining information regarding the significance and nature of population.

ELEMENTS OF SAMPLING

The sampling process consist of following three elements:

i. Sample selection,

ii. Collection of information,

iii. Drawing the inferences about the population from the selected sample.

All the three elements are interwoven and has an impact on each other. Sampling follows a set of rules while selecting a sample and the estimation is not independent but guided by way in which sample is selected.

ESSENTIALS OF SAMPLING

In order to get worthwhile inference from the selected sample, it is essential that the sample should possess the following essentials:

- The selected sample should be representative of the universe. In order to ensure this, essential sample should be selected at random.
- The size of the sample should be adequate or large so that it can represent the characteristics of population.
- The nature of units of universe and that of the sample are homogenous.
- The items of selected sample are independent of each other and have equal chance of being selected.

IMPORTANT TERMS TO BE REMEMBERED

The number of units in the sample is known as the *Sample size*. The value obtained from the sample study such as averages is known as *Statistic*. Statistics are functions of sample observation. When the values such as variance, skewness, kurtosis, correlation coefficient, etc., are obtained for the population, it is known as Parameters. Parameters are functions of population.

THEORETICAL BASIS OF SAMPLING

The behavior of the mass phenomena or universe is predicted and generalized on the basis of the study of sample because there is no statistical population whose elements vary from each other without limit. For example, though rice or wheat vary in color, proteins, contents weight, etc., to a limited extent, they are still identified as rice or wheat. Similarly, though there is diversity in the population, the entire population has similar characteristics with limited variations. So it is possible to select a relatively unbiased random sample which can fairly represent the traits of the population.

The theory of sampling is based on the two important laws. They are:

1. Law of Statistical Regularity, and
2. Law of Inertia of Large Numbers.

Law of Statistical Regularity

The law of statistical regularity owes its origin to the mathematical theory of probability. According to Conner, "The law of statistical regularity lays down that a group of objects chosen at random from a larger group tends to possess the characteristics of that large group (universe)." In the words of Kings: "The law of statistical regularity lays down that a moderately large number of items choose at random from a large group almost sure on the average to possess the characteristics of the large group." In simple words, the law states that if the sample is drawn from the population at random, is likely to have the same characteristics as that of the population. So the two important points are:

i. **Large Sample Size:** As the sample size increases, the sample is likely to reveal the characteristics similar to population and provides reliable estimates of the population parameter.

ii. **Random Selection:** The sample from the population is to be selected at random. By random selection, we mean a selection where each and every item of the population has an equal chance of being selected in the sample. An randomly selected sample would be representative of the population.

If a selected sample satisfies the above two conditions, it will depict fairly and accurately the characteristics of population and will help in drawing valid inferences about the population. For example, if we want to study the average height of a student at ICFAI University, then it is not necessary to study the height of each and every student. We can choose few students at random from every center or college, measure their height and draw an inference of the average height of students at ICFAI university.

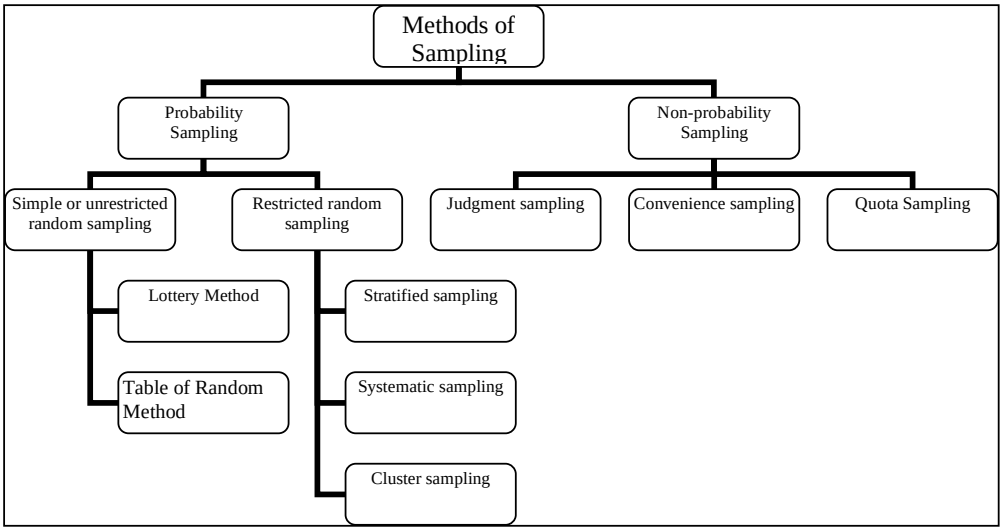
Law of Inertia of Large Numbers

The law of inertia of large numbers is corollary and immediate deduction from the principle of statistical regularity. This law is of great significance in the theory of sampling. It states that "The *other things being equal, larger the sample size, more reliable and accurate the results tends to be."* This is because, when compared to small size, the large numbers are more stable. When the sample size is large, the difference in aggregate result is insignificant because variation in one part of the universe in neutralized by variation in bigger part of the population. In the words of A. L. Bowley "*Great numbers and averages resulting from them, such as we always obtain in measuring social phenomenon have a great inertia."* So it cannot be inferred that the large numbers have no variations. The large numbers do exhibit the variations but are of small magnitude and intensity. For example, if a coin is tossed for 20 times then it is difficult to talk about the proportion of heads and tails. Because the experiment is carried on for small number of times, it is possible that we may not get 10 heads and 10 tails. The result can be a combination of 15 heads and 5 tails, or 16 tails and 4 heads or 12 heads and 8 tails and so on. But if the same experiment is carried on for 1000 or 5000 times, the chance of getting equal heads and tails will be very high, because the experiment is repeated for large number of times, the variation in one direction is compensated by the variation in other direction.

METHODS OF SAMPLING

Based on the laws, the selection of a sample is undertaken. This process involves different methods. These can be studied from the following figure:

Figure 1



As observed in the above figure, the sampling methods are divided into two heads. They are:

1. Probability Sampling or Random Sampling.
2. Non-probability Sampling or Non-random Sampling.

Probability Sampling

When every item in the universe has a known chance of being selected in the sample, such a method is known as Probability Sampling. It is a scientific technique of drawing sample from the population according to the law of chance. According to the law of chance, each unit or item of the population has pre-assigned probability of being chosen in the sample. This implies that the sample items are selected at random and are independent of the person making the study.

According to Simpson and Kafka, "random samples are characterized by the way in which they are selected. Randomness is not used in the sense of haphazard or hit or miss."

Advantages

The probability sampling has the following advantages:

i. Probability sampling does not require the detailed information about the population for its effectiveness.

ii. The estimates provided by probability sampling are unbiased and have measurable precision.

iii. Evaluation of relative efficiency of various sampling design is possible.

Limitations

The followings are the limitations or disadvantage of probability sampling due to which non-probability sampling is not widely used in practice.

i. High level of skill and experience is required for using probability sampling,

ii. Planning and execution of probability sampling requires lot of time,

iii. It is costly when compared to non-probability sampling.

METHODS OF PROBABILITY SAMPLING

1. Simple or Unrestricted Random Sampling
2. Restricted Random Sampling:
   a. Stratified sampling
   b. Systematic sampling
   c. Cluster sampling.

Simple or Unrestricted Random Sampling

Simple or Unrestricted random sampling is a sampling technique in which the sample drawn is such that each and every unit in the population has an equal chance of being included in the sample. In simple random sampling, the item selected in sample is just a matter of chance. As said earlier, the word random does not mean haphazard or hit or miss but it implies that only the chance, which determines the item that are to be included in the sample.

The simple random sample may be without replacement and with replacement. When the unit or item selected in the first draw is not replaced in the population before making the second draw, such a sampling is known as Simple random sampling without replacement. When the item selected in the first draw is replaced in the population before making the next draw, such a sampling plan is known as Simple random sampling with replacement.

According to Chou, the sample is a simple random sample when 'n' the sample size is drawn from a population with N element if any of the following conditions are true:

- All the N items in the population have an equal chance of being included in the sample and the sample size 'n' is drawn independently of one another.
- At every selection, the remaining items in the population have equal chance of being drawn. If the sampling is with replacement, then the probability of each item being drawn at each selection is 1/N. If the sampling is without replacement, then the probability of selection of each remaining item in the population in first draw, is 1/N, in the second draw it is 1/(N – 1), in third draw it is 1(N – 2) and so on.
- All the given sample size of 'n' are equally likely to be selected.

Selection of Simple Random Sample: A random sample may be selected by –

- **Lottery Method:** This is the simple method and independent of the properties of the population. Under this method, various units are numbered on small slips or cards and shuffled thoroughly in a drum from which the desired sample size is drawn. It is the most reliable methods of random sampling.
- **Table of Random Method:** The limitations of lottery method viz., bulkiness and consumption of more time made it significant to use this method. Under this method different tables are constructed and the numbers are assigned for each unit of population of the size "N" in the table. These numbers range between 0 to 9 and appear with the same frequency independent of other. It involves the following steps:
  - Identify N units in the population with the numbers 1 to N.
  - Select at random any page of the random table number and pick up the numbers in any row, column or diagonal at random.
  - The population units corresponding to the numbers selected in second step constitute the random sample.

Merits

i. There is no possibility of personal bias or judgment of the investigator affecting the result as the sample selected depends entirely on chance.

ii. The selection of sample according to this method is considered to be more representative of the universe when compared to judgment sampling. Larger the size of sample, more it becomes representative of the population.

iii. It is possible to ascertain the efficiency and accuracy of the estimate because the sampling error follows the principle of chance.

iv. Random sampling provides the reliable and maximum information at the least cost and therefore leads to saving in time, money and labor.

Limitations

i. Simple random sampling necessitates complete and up-to-date information of the population from which the sample is to be drawn. But in practice, it is not possible to obtain the complete information, which restricts the use of sampling design in economic and business data.

ii. For ensuring the statistical reliability and accuracy, the size of sample required in simple random sampling should be large when compared to stratified sampling.

iii. It is time consuming and costly when we have to collect the requisite data for the survey which is fairly large and scattered widely and geographically.

iv. This method may not reflect the true characteristics of the population or may not be representative of population if the sample is not large.

v. It is uneconomical and time consuming if the population is large.

vi. It may produce most non-random looking result.

Restricted Random Sampling

Stratified Sampling

Stratified random sampling is a random method, which uses the available information relating to population for designing a more efficient sample. It involves the following steps:

- The given population to be sampled is sub-divided into number of sub-groups or sub-population known as Strata. The units of each stratum are homogeneous and differ as widely as possible.
- Draw a simple random sample independently from each of the strata.

  The stratified sampling differs from simple random sampling, i.e., in simple random sampling, the items of sample are chosen at random from the entire universe whereas in stratified random sampling, the items constituting sample are chosen from the stratum.

Selection of stratified random sampling involves the following steps:

a. The population is divided in strata on the basis of known variable, correlated with variable of interest and possesses the information on each element of universe. A constructed strata should minimize the difference among the sampling units within strata and maximize difference among the strata.

b. The number of strata to be constructed should be feasible otherwise the cost of construction of strata may out run the benefit.

c. The decision of sample size within strata is taken either on the basis of proportional allocation or disproportional allocation.

The stratified sampling may be proportional stratified sampling or disproportional stratified sampling. When the number of items selected from each stratum is independent of its size it is called as disproportional stratified sampling. Whereas in proportional stratified sampling, the number of items drawn from each strata are proportional to its size in the strata.

Merits

The following are the advantages of stratified random sampling:

i. The stratified random sample is more representative. Because we first divide the population into various strata and then draw a sample from each stratum. So, the possibility of any important group of population being ignored is completely eliminated.

ii. It provides greater accuracy and more efficient estimates when compared to simple random sampling as it reduces the variability within each stratum. The accuracy is maximum if each stratum consist of uniform or homogenous item.

iii. When compared to simple random sampling, stratified sampling has greater geographical concentration i.e., the units from various strata are selected in such a way that they are localized in one geographical area. This ultimately results in reduction of cost, time, expenses of interviewing and supervision of field work.

Demerits

The following are the limitations of stratified random sampling:

i. Division of population into different strata requires utmost care because the success of stratified sampling depends on it. In order to get reliable result, each stratum should have homogeneous items.

ii. Selection of items from stratum at random may become difficult in the absence of skilled sampling supervisor.

iii. The cost of stratified sampling will be high when compare to simple random sampling as the samples are geographically distributed.

iv. If the weights assigned to different strata in disproportional stratified sampling are faulty, then the resulting sample may not represent the population and might give biased results.

Systematic Sampling

In systematic random sampling, the first unit is selected at random and then the additional units are selected automatically in a definite sequence at evenly spaced intervals from one another until the sample is formed. It is the slight variation of simple random sampling. This technique of drawing the sample is popularly used and recommended in those cases, where the complete and up-to-date list of population from which the sample is to be drawn is available. The list is prepared and arranged in systematical order such as alphabetical, geographical, chronological, numerical, etc. The items or the sampling unit 'N' are arranged in systematic order and serially numbered from 1 to N and we have to draw 'n' sample size from it. The lottery method is followed for selecting the first item at random and the subsequent items are selected by taking the kth item from the list. Where 'k' refers to the sample interval or sampling ratio i.e., the ratio of population to the size of sample.

$$N = nk \Rightarrow k = \frac{N}{n}$$

Where, n = sample size, k = the sample interval, and N = Size of universe.

We follow approximation procedure if we get a fractional value while calculating k i.e., if the fractional value is less than 0.5, then it will be omitted and if it more than 0.5 then it is taken as 1. If the fractional value is exactly 0.5, it will be taken as 1 if the number is odd and will be omitted if the number is even.

Merits

i. Systematic sampling is simple, easy to operate and convenient to adopt.

ii. The time, labor and work involved is considerably less when compared to simple or stratified random sampling.

iii. The results are satisfactory, provided utmost care is taken for avoiding the periodic feature associated with the sampling ratio.

iv. When the frame is complete and up-to-date and the units are serially arranged in a random order then systematic sampling is more efficient than stratified sampling.

Limitations

i. The first drawback of this method is requirement of complete and up-to-date frame of sampling unit and its arrangement randomly. If the requirement is not fulfilled then the method may not work well.

ii. Systematic sampling may become less representative or may give bias results if there are periodicities in the population and the sampling interval is equal to period.

Cluster Sampling

In cluster sampling, primary, secondary and final units are selected randomly from a given population or stratum. The sampling process is carried out in various stages. Under this method, first stage units are sampled by some suitable method, then the samples of second stage units are selected from selected first stage unit again by some suitable method. As per the requirements, further stages are added in the similar way. For this reason, this method is also known as Multi-stage sampling.

Merits

i. It introduces flexibility which is lacking in other methods.

ii. It utilizes the existing divisions and sub-divisions of the population at various stages, thereby covers a large area and at the same time permits the work to be concentrated.

Limitations

The only drawback of this method is that the accuracy is less in cluster sampling when compared to any other method.

Non-Probability Sampling

When the process of selecting the sample is without the use of randomization, it is known as Non-Probability or Non-random sampling. In non-random sampling, samples are selected on the basis other than the probability consideration.

METHODS OF NON-PROBABILITY SAMPLING

The non-probability sampling methods are:

i. Judgment Sampling,

ii. Convenience Sampling,

iii. Quota Sampling.

Judgment Sampling

In this method, the selection of sample depends on the judgment of the investigator exclusively. It includes those items, which truly represents the universe with regard to the characteristics under investigation. For example, for studying the spending habits of the students, a investigator may select five or ten students from a class of hundred students who in the opinion of the investigator are representatives of the class.

Merits

i. When the numbers of sampling units are small in the universe.

ii. For studying the unknown traits of the population from some of the known characteristics. Then stratifying the population as per these characteristics and selecting a unit from each stratum on the judgment basis.

iii. When time is a limiting factor as in the case of business and public policy. In such case, probability sampling cannot be employed and judgment sampling is the only the practical method for arriving at solution for such problems.

Limitations

i. The method is not scientific because investigator's personal prejudice or bias may affect the sample unit.

ii. The success of this method depends on the excellence of the investigator. As such there is no objective way for evaluating the reliability of sampling result and also for determining the size of sampling error.

Convenience Sampling

When the sample is obtained by selecting a convenient population unit, the sample is known as Convenience sample. The convenience sampling is also called the *chunk*. A chunk is a part of population being investigated that is selected neither by probability nor by judgment but by convenience. The examples of convenience sampling are samples obtained from telephone directories, automobile registration, etc. Personal bias effects the convenience sampling because the investigator will choose a sample that is convenient to him. So, the convenience sample selected will not truly represent the population. Therefore, convenience sampling is used for pilot studies. Before deciding on the sampling design, we test the questions and obtain the preliminary information.

Quota Sampling

The most commonly used sampling technique in non-probability sampling is Quota sampling. It is a kind of judgment sampling. According to specific characteristics, quotas are set up. The characteristics may be income group, age group, professionals, etc. Then the interviewer is asked to interview certain number of persons from a given quota. Within this quota, sample is selected on the personal judgment of the interviewer. For example, in conducting a TV serial survey, the interviewer may be asked to interview 1000 persons living in certain locality and out of every 100 persons interviewed, 50% should be housewives, 25% should be old-aged people and 25% should consist of teenagers and children. The interviewer has freedom of selecting the people from this quota. In quota sampling, the cost per person is small but opportunities for the personal bias invalidating the result is very high.

Quota sampling is similar to stratified random sampling because in both the methods, the population is divided into parts and then the sample is selected. But it differs from stratified sampling in the sense that sample are chosen at random from each stratum whereas in quota sampling, the sample is not selected at random. Quota sampling is used widely in public opinion studies. This method will give satisfactory result if the interviewer is trained properly.

SIZE OF SAMPLE

An important decision in the sampling technique is about the size of the sample. The term size of sample refers to the number of sampling units selected for investigation from the population. The experts with respect to size of sample express different opinions. Some suggested it should be 5% of the population; other argued it should be 10% or 25% of the population. However it should be noted that size is not representative of the population. A well-selected sample though smaller in size is always superior to the badly selected large sample. So, the size of sample should neither be too large nor too small but it should be optimum. According to *Parten*, an Optimum size is one that fulfils the requirements of efficiency, representativeness, reliability and flexibility.

The factors that should be borne in mind while deciding the sample size are:

- **The Size of the Population:** The general rule is that the larger the size of population or universe, the bigger should be the size of sample.
- **Availability of Resources:** A larger sample size could be considered if the resources at disposal are abundant. Usually, non-availability of resource constitutes constraint on the size of sample.
- **Desired Degree of Accuracy and Precision:** The larger the sample size, greater is the degree of accuracy. But it does not mean that larger sample ensures greater accuracy. A small sample selected by scientific method may ensure better result and greater accuracy then a large sample selected by inexperienced investigator.
- **Homogeneity and Heterogeneity of the Universe:** A small sample will serve the purpose if the universe consists of homogeneous units and a large sample is needed when the universe consists of heterogeneous units.
- **Nature of Study Undertaken:** A small sample may be suitable if the study undertaken is intensive and continuous or repeated. A large sample is needed when the study is extensive in nature.
- **Sampling Method Adopted:** The method adopted for selecting an sampling also influences the size of sample.
- **Nature and Response of the Respondents:** When the number of respondents are large and will not co-operate, then a large sample is preferred.

Determination of Size of the Sample

Depending on the availability of information, number of formulas are devised for determining the size of sample. The popularly used formula is

$$n = \left[\frac{Z\sigma}{d}\right]^2$$

Where,

n – Size of sample.

Z – Critical value of the standard normal variate at specified level of significance, 1% or 5%

σ – Standard deviation of the population.

d – Difference between population mean and sample mean.

MERITS AND LIMITATIONS OF SAMPLING

Merits

The following are the merits of sampling technique over the census or complete enumeration survey:

i. **Less Time:** Since sample study, inspect and examine only a part of the population or universe, there is considerable amount of saving in time and labor when the sampling technique is followed. Time is saved not only in conducting enquiry, collecting data but also in processing, editing and analyzing it. Thus, the sample provide timely, quickly and urgently needed data when compare to complete enumeration survey.

ii. **Economy:** The sample technique is more economical than the complete enumeration survey. Though in sampling method, the effort and cost involved in collecting information per unit is greater than census method, but the total financial burden is less than the census method. Because in sample we study only a part of population. This is particularly advantageous in conducting socio-economic surveys in developing countries and in under developed countries which cannot afford census method due to lack of adequate financial resources.

iii. **Reliable Results:** In census method, the sampling errors are totally absent but the existence of non-sampling error does not give 100% accurate result. Whereas sampling method involves certain inaccuracies due to existence of both sampling and non-sampling error. Inspite of this drawback sampling survey gives results that are more reliable than census method. Because the extent of sampling error can be determined and the other types of errors are more serious in complete enumeration method, which can be effectively controlled and minimized in sampling technique.

iv. **Detailed Information:** In complete census method, there is possibility of obtaining the detailed information as each and every part of the universe is enumerated. But in practice, detailed and exhaustive information can be collected in sampling method as the sampling technique saves time, labor and money. The sampling method is more readily available when compare to census method. Because highly trained personnel and sophisticated equipments are needed for collecting, processing and analyzing the data in census method.

v. **Infinite or Hypothetical Population:** In case of large population, sampling is the best technique for estimating the parameters of the population. For example, the number of lions, tigers in a thickly densed forest can be estimated only by sampling method. Similarly in hypothetical population such as tossing of coin, the only scientific technique for estimating the population parameter is sampling.

vi. In certain cases, such as destructive testing the sampling technique is only the practical mean. Destructive testing, such as estimating the average life of bulbs, breaking strength of chalks manufactured by a factory, etc.

vii. Lastly, sometime sampling technique is used for judging the accuracy of the information obtained by complete enumeration method.

Limitations

Sampling technique is not free from limitations. Some of the difficulties involved in sampling technique are as follows:

i. If the sample survey are not properly planned and carefully executed, the results obtained may be misleading and inaccurate. In the words of Stephen:

"*Samples are like medicines. They are harmful when they are taken carelessly and without the knowledge of their effects*".

ii. Sampling requires services of qualified, skilled experts, effective supervision and sophisticated equipments and statistical techniques for collecting, analyzing the sampling data. In the absence of these perquisites, the information provided by sample cannot be relied upon.

iii. Complicated sampling techniques may sometimes need more time, labor and money than the complete census method. Because, the size of sample is large proportion of the total population, size and complicated weight procedure is employed.

iv. When the information for each and every unit of population is needed, we cannot employ sampling technique and the purpose is best served by the complete enumeration survey.

SAMPLING AND NON-SAMPLING ERRORS

The term error denotes the difference between the true or actual value and the estimated or approximated value. In statistic, error refers to discrepancies or the difference between the actual value of population parameter and its estimate provided by an appropriate statistical device. In statistical investigation, the inaccuracies of error are broadly classified as Sampling and Non-sampling error. It is necessary to understand the role of these errors in census and sample surveys in order to know the importance of sample. Sampling errors are the errors that arise due to drawing of inferences about the population on the basis of sample. This error does not exist in complete enumeration survey. The errors that arise in ascertainment and processing of data is known as Non-sampling error. This error exists in both sample survey and census method.

Sampling Error

The results obtained from the sample study may not be exactly similar to the actual value of population even if utmost care is taken. The reason being that in sample survey, only a part or small portion is studied and not the whole population. So, the results are bound to differ from the census results and have certain amount of errors attributed to sampling fluctuation. Such errors are known as Sampling errors. The sampling error arise due to the fact that only the subset of the population is studied for estimating the population parameter and drawing inferences about the universe. Thus, sampling errors are present in sample survey and absent in complete census method. However, these errors can be controlled. There are two types of sampling error:

BIASED ERROR

The error arising from any bias in selection, estimation is known as biased errors. The biased errors remains constant i.e., it does not decrease as the size of the sample increases. Therefore, such errors are also known as cumulative or non-compensating error.

UNBIASED ERROR

The errors arising due to chance difference between the population member selected in sample and those not selected are known as unbiased error. The unbiased or random error decreases as the size of the sample increases. Therefore, such errors are known as non-cumulative or compensating error.

The biased error arises due to following causes:

i. **Faulty selection of sample** may give rise to bias in following way – the investigator deliberately selects a sample to obtain certain result, bias in selection of a random sample consciously or unconsciously, investigator substituting an item in the place of one chosen, when all the items to be covered are not covered, etc.

ii. **Faultly collection of data:** An consistent error in measurement will give rise to bias whether the measurement is carried on sample or on the population.

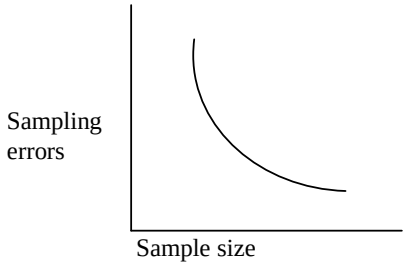
iii. **Faultly Method of analysis:** Not only that the faulty selection of sample and collection of data give rise to bias but the bias may also arise due to faulty method of analysis.

The existence of bias affect the objective conclusion. The bias can be avoided if the sample is selected at random or at random subject to restriction.

Methods of Reducing Sampling Error

Once the absence of bias is ensured, attention must be given for reducing the random sampling error to the minimum and attaining the desired accuracy. The accuracy of the sample can be increased by increasing the size of sample. The sampling errors are decreased by increasing the size of sample and the decrease inversely proportional to the square root of the size of sample as shown in the figure 2 given below:

Figure 2



From the above figure it is clear that an increase in the sample size initially leads to decrease in the sampling errors but after certain stage it becomes marginal. It implies that greater efforts are needed for reducing the sampling errors in later stage when compared to initial stages.

The non-sampling errors are more in complete census method than in the sample survey method. It can be reduced by using better organization and trained personnels at field and tabulation stage. The non-sampling errors increase with the increase in the size of sample. Sometimes, the non-sampling errors in complete census method may be more than the total sampling and non-sampling error of the sample survey.

Non-Sampling Error

When the complete enumeration method is used, one can expect that the data is free from errors. But in practice it is not so. Errors, which are not attributed to chance and arise due to factors, which are within the human control, are termed as Non-sampling error. In the other words, the errors arising at any stage of enquiry such as planning and execution of survey, collection and analysis of data are Non-sampling error. These errors are present in both complete enumeration and sample survey method. The Non-sampling error increases as the size of sample increases; they often arise in census method rather than in the sampling method. The Non-sampling errors arise due to following factors:

- Faulty planning and definition of population or the sampling unit.
- Errors due to non-responses.
- Inappropriate statistical unit.
- Imperfect questionnaires resulting in incomplete information.
- Defective interviewing methods and questions.

- Lack of trained, experience investigators, and inadequate inspection of primary staff.
- Errors committed during presentation and printing of results.

Control of Non-sampling Errors

The non-sampling errors deserves greater attention as they are large in number when compare to sampling error. The sampling errors are reduced as the sample size increases but the non-sampling errors increases as the sample size increases. These errors should be reduced to a level so that their presence does not affect the final result.

Reliability of Sample

The reliability of sample depends upon the following tests:

- From the same universe, take more sample and compare their results. If the results are same, then the sample is reliable.
- If the measurement of population is similar to the measurement of the sample then the sample is reliable.
- If the results of the sub-samples taken from the sample are similar then the sample is said to be reliable.

SUMMARY

- The study of entire population is known as census or complete enumeration method.
- A sample is a subset of population. It is that past of the universe that is selected for the purpose of investigation.
- The study of the sample is known as sampling there are two methods of sampling. They are: (i) Probability/Random sampling, and (ii) Non-probability or non-random sampling
- Probability sampling methods are those in which every item of populations has a chance of being chosen in the sample.
- Non-probability sampling methods are those in which every item of universe does not have chance of being included in the sample.
- The error that arise due to drawing of influence about the population on the basis of few observation is known as sampling error. The error's that arise during collection and processing of data are known as non-sampling error. These errors arise in both sample survey and census method.

Chapter IV

Classification and Tabulation of Data

After reading this chapter, you will be conversant with:

- Meaning and Objectives of Classification
- Types of Classification
- Tabulation of Data
- Parts of Table
- Rules of Tabulation
- Types of Tables
- Table Review
- Frequency Distributions

Introduction

In previous chapters, we have learnt the process and various methods of collection of data. Usually this data is huge and contained in schedules and questionnaires, such data is known as raw data. After the collection of data the next important step is to classify and tabulate the collected information or to rearrange them into new groups, if the data is collected from the published statistics. This aspect of statistical enquiry is termed as organization of data. The data should be presented in comprehensible condensed form so that the important characteristics of the data are highlighted and also facilitates further comparison, processing and interpretations. Data processing operations include the following:

- Editing,
- Coding,
- Classification, and
- Tabulation.

The presentation of data is classified into two categories. They are:

- Tabular Presentation,
- Diagrammatic or Graphic Presentation.

The tabular presentation is the next step after the collection of data. But, systematic arrangement of raw data into homogeneous classes is must before proceeding to tabulation of data. It is necessary for sorting out the necessary relevant and significant information from irrelevant and insignificant data, which is done through classification of data. Thus, the primary or preliminary step in tabulation is classification of data because the similar items are to be brought together before presenting the data in the form of tables. The present chapter deals with the classification and tabulation of data and the other forms of presentation are discussed in subsequent chapters.

MEANING AND OBJECTIVES OF CLASSIFICATION

Classification is the first step towards the further processing after the collection, editing and coding of the data. Classification, technically, is nothing but grouping of related or similar items into classes according to their similar features. It is one of the tools in the processing of data.

According to Secrist “Classification is the process of arranging data into sequences and groups according to their common characteristics, or separating them into different but related parts.”

Thus, classification is nothing but arrangement of data into classes determined on the basis of nature, objective and scope of the enquiry. When the data are sorted on one basis of classification and then again on another basis simultaneously, such classification is known as cross-classification. This process can be repeated for many times as long as classification is possible. Classification is somewhat similar to sorting of letters in a post-office. In post office the collected letters are sorted into different lots on geographical basis and are put into separate bags containing letters with a common feature. Similarly, the number of students registered in ICFAI University during the academic year 2008-2009 can be classified on the basis of different criteria such as age, the state to which they belong, sex, religion and the faculties to which they have joined. This data can be classified in number of ways into different groups on the basis of some recognizable characteristics. This fact is known as basis or criteria for classification. Thus, the process with which the information is obtained in summary form is called as classification of data. Based on the above description, the following points are noteworthy:

- The collected data must be segregated and distributed into different groups or classes.

- The segregation and distribution must be based on the common characteristics, i.e., the data possessing similar characteristics must be brought under one group or class.
- The groups can be actual. For example: rich and poor, rural and urban etc.
- In the end, it is an important aspect of classification that it should bring out the uniformity in the diverse data.

Objectives of Classification

The basic and principal objectives of classification of data are:

- The voluminous and huge raw data is condensed in such a way that the similarities and dissimilarities of the data are apprehended readily.
- To facilitate meaningful comparison of the data on the criteria of classification.
- To point out the important and significant features/characters contained in the data.
- To classify the data on two or more criteria for studying the relationship between these two criteria.
- To gather the relevant and important information while dropping the irrelevant and unnecessary elements.
- To facilitate the statistical treatment of the voluminous heterogeneous data into relatively homogeneous groups for further processing such as tabulation, analysis and interpretation of the data.

Box 1
Rules of Classification

The following are the general rules or principles that should be followed in the process of classifying the data:

1. The classification must be exhaustive i.e., every item in the data should be covered in a particular group or class during classification.
2. The overlapping of groups or classes should be avoided i.e., the contents of segregation or distribution should be mutually exclusive.
3. The classification should comply with the objective statistical investigation.
4. To avoid difficulties in drawing inferences the basic objective of classification should be maintained through out the process. Thus, maintaining stability is one of the principles of classification.
5. As discussed above, the items of classification should be homogenous and have similar characteristics.
6. A good classification should be flexible viz., it should be adaptable to the changing circumstances.

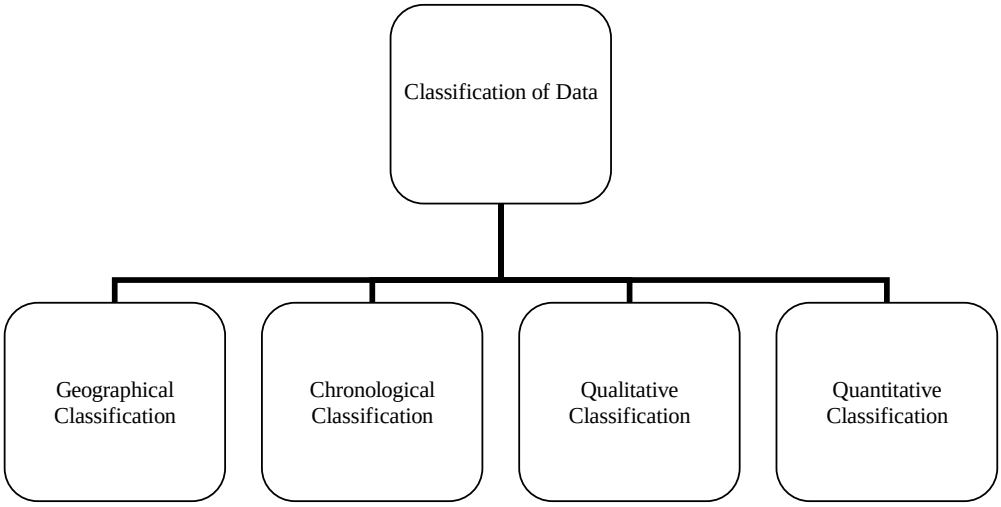
Source: Statistics (Theory, Methods and Applications) by D.C.Sancheti and V.K. Kapoor.

TYPES OF CLASSIFICATION

The classification of data depends on the objectives and the purpose of the enquiry. Thus, the data can be classified on the basis of following criteria:

- Geographical i.e. Area-wise or regional. ex. States, Cities, Districts, Towns.
- Chronological i.e. on the basis of occurrence of time.
- Qualitative i.e. according to some character or attributes.
- Quantitative i.e. according to terms of magnitudes or numerical values.

Figure 1



Thus, the above classification can be dealt in detail as under:

Geographical Classification

Geographical classification is classification of data on the basis of geographical or locational differences between the various items of data such as states, cities, regions, zones, areas, etc. For example, the production of agricultural output per hectare in India may be presented in the form of state wise allocation as shown below:

State-wise Estimates of Agricultural Output in India

Name of States	Output per hectare
Andhra Pradesh	125
Bihar	150
Gujarat	145
Maharashtra	175
Punjab	130

The geographical classification are usually listed either in alphabetical order or according to value for stressing the importance of area or region.

Chronological Classification

Chronological classification is a classification in which data are classified on the basis of time. In other words, when the data is presented over a period of time, then the classification adopted is known as chronological classification. For example, population of India over different years, industrial production over a period of time, etc.

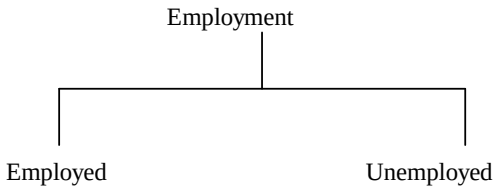
Population of India from 1951 to 2001 (In Crore)

Year	Population
1951	36.11
1961	43.9
1971	54.8
1981	68.3
1991	84.6
2001	102

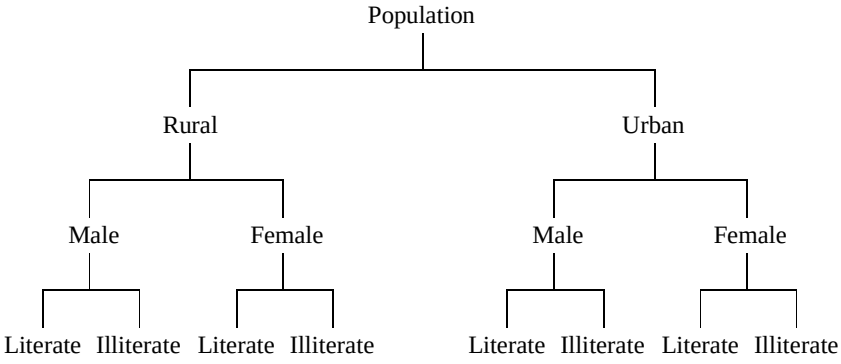
Time series data is usually classified chronologically, which is used frequently in Economics and Statistics.

Qualitative Classification

In qualitative classification, data is classified on the basis of qualitative phenomenon such as attributes or quality, which cannot be expressed in quantitative measurements such as sex, beauty, honesty, color of hair, literacy, intelligence, occupation, employment, etc. Such classification is also known as Descriptive classification or classification as to attributes. The important feature of qualitative classification is that the attribute under study cannot be measured and the data is classified on the basis of presence or absence of the attributes in the given unit of population under study. When the data is classified into two classes with respect to presence and absence of character or attribute, such a classification is termed as **Simple or Dichotomous Qualitative Classification**. For example, when the population is divided on the basis of employment, one can find out how many are employed and how many are not employed.



When the two classes of population are further divided on the basis of some other attributes so as to form several classes or groups, such a classification is known as **Manifold classification**. The example of manifold is given below:



Quantitative Classification

When the data is classified according to some characteristics or phenomenon, which are capable of quantitative measurement, such as weight, height, prices, profits, sales production, expenditure, etc. This quantitative phenomenon under study is known as variable and this classification is also termed as classification by variables. For example, the students of a college classified according to their marks are as follows:

Marks	No.of Students
0-10	6
10-20	14
20-30	8
30-40	10
40-50	18
50-60	6
60-70	12

The explanation regarding variables is discussed in later pages of the chapter.

TABULATION OF DATA

One of the simplest devices for summarizing the data and for the systematic presentation of information contained in data in the meaningful manner is known as statistical table. A table is a systematic arrangement of statistical data in rows and columns. The horizontal arrangement is known as rows and the vertical arrangement is called columns. The act of preparing a table is called tabulation. Tabulation simplifies the presentation and facilitates comparisons and the work of further statistical analysis. It is an ingenious device of presenting the data in a comprehensible and condensed form by providing the maximum information in the minimum possible space without affecting the quality and usefulness of the data.

Professor Bowley defined tabulation as "the intermediate process between the accumulation of data in whatever form they are obtained and the final reasoned account of the results shown by the statistics."

It is an intermediate process between data collection and statistical analysis. It is the final stage in collection of data and provides scope for further statistical analysis and interpretations.

Tabulation and classification go together in statistical analysis. Classification is the first step in tabulation. We have to classify the data according to common characteristics before putting the data in tabular form. It is only after the classification the data is presented in the forms of rows and columns so as to have clear understanding of relationship between them.

Significance of Tabulation

The importance of tabulation is summarized as follows:

i. Tabulation simplifies the huge and complex data by avoiding the irrelevant and repeated facts. It presents the data systematically in rows and columns so as to enable the reader in having the clear understanding of it.

ii. Tabulation facilitates comparison by dividing the data into several parts with their totals and sub totals. This helps in studying the relationship between the different parts of the table.

iii. Tabulation gives identity to the data by arranging it in a table with title and numbers. This facilitates the interpretation of the problem.

PARTS OF TABLE

The parts of the table vary from problem to problem depending upon the given data, its nature and the investigation purpose. However, the following are the main part of a good statistical table:

Table Number: Each table should be numbered in logical sequence for identification and future reference. The number can be placed either on the top or at the bottom or in the center at the top.

Table Title: Every table should have a suitable title, which appears on the top of the table or below or next to table number. A title is a brief and concise description of the contents of the table and is self-explanatory without sacrificing the clarity. The title should describe the nature of the data, place, time and source of the data. The title should be in the form of series of phrases and prominently lettered.

Caption: The heading or description of the vertical columns is known as caption. It is written in the middle of the top of the column representing what it explains. It should be brief, clearly defined, concise and self-explanatory. There may be sub-heads under a column. The caption may be given as a head note along with the title if the same units are used for all the entries of the table. If the different columns are expressed in different units, the corresponding units are indicated in column in smaller letters.

Stubs: The designation of horizontal rows or the heading of the rows is termed as stubs. They perform the same function for the horizontal rows as the caption perform for the vertical columns of the number in the table. Stub is placed on the extreme left of the table. Sub-stubs are formed for rows with similar classification and may be grouped under a common heading to avoid repetitions.

Body of the Table: Body of the table contains the numerical information of data arranged according to the description given in the caption and stubs form. It forms an important and vital part of the table.

Headnote: A brief explanatory statement that applied to all or major part of the table is known as Headnote. It is placed below the title at the centered point and enclosed in the bracket. It explains the points relating to the entire table, which are not explained by the title, caption and stub are considered as supplement to the title. For example, we write the headnote for unit of measurement such as in rupees, in million etc.

Footnote: Footnote is used when the features of the table are not adequately explained by the title, caption and stub or when additional information is needed for complete description. As the name indicates, footnote are placed below the body of the table at its bottom. It is identified by the symbols such as *,**,***, etc.

Source Note: When the secondary data is used, we use source note. It is given at the bottom of the table below the footnote. It helps the user in deciding about the reliability of data.

The following figure gives the format of a table:

(Headnote if any)

Stub heading	Caption				Total (rows)
	Sub-Head		Sub-Head		
	Column Head	Column Head	Column Head	Column Head	
		Body			
Total (Columns)					

Footnote:

Source Note:

RULES OF TABULATION

There are no hard and fast rules for tabulation of data because it depends on the data given and the purpose or the objective of survey or enquiry. In order to prepare good table one should have clear idea about the facts to be presented, the points which are to be emphasized and practical experience in the preparation of table. Thus, construction of a good table is an art, which can be obtained through skill, experience and expertise of the tabulator. Prof. Bowley rightly stated that "*In collection and tabulation common sense is the chief requisite and experience is the chief teacher.*"

However, while tabulating the data following rules should be considered:

i. A table should have more rows and the columns should be less than the rows. It should provide space for references.

ii. A table should be simple to understand and free from ambiguities and overlapping. It should be complete and self-explanatory.

iii. The captions and stubs of the table should be arranged systematically. The items of the table are arranged as per the nature of the data. It may be arranged either in alphabetical, chronological, geographical, conventional manner or according to the size of the data.

iv. The units of measurement such as in Rupees, weights in pounds, height in inches, etc., are to be clearly specified in the table.

v. Fractional figures should be rounded off to the nearest and a footnote is given to that effect.

vi. If the data has many details then it should not be overloaded in one table. Instead number of tables may be prepared to show the distinct character.

vii. The table should be arranged logically and the related items should be placed to the extent possible in the same group. The rows and columns should be numbered for identification.

viii. Figure in table conveys meaningful information if expressed in percentages or ratios. If we want to compute and show percentages and ratios, then additional column should be inserted in the table.

ix. If the information is not available then show it as N.A. and not zero. Zero should be written only when the quantity is zero.

x. Avoid abbreviations especially in titles and headings.

xi. Ditto marks should not be used for repeated figure. Because it may be mistaken as '11'. Write the repeated figure each time.

xii. The important feature of tabulation is clarity. So the tabulator should be explicit and should not use the expression 'etc'.

xiii. Since it is difficult to follow all the above guidelines while preparing a table, J.C. Capt accordingly suggested that a tabulator should follow two rules. The two rules are – (1) While planning the table use common sense, and (2) View the proposed table from the stand point of the user.

TYPES OF TABLES

Statistical table is broadly classified into two categories:

1. Simple and Complex Table.
2. General Purpose Table and Special Purpose Table.

Simple and Complex Table

This classification is based on the characteristics studied or on the basis of coverage. When the data is classified on the basis of single characteristics, it is the case of simple table. It is also known as one-way table. On the other hand, when the data is classified into different classes on the basis of two or more character simultaneously, it is complex table. When in the table, two characters are shown, such table is known as two-way table or double tabulation. When the table shows three characters, then the tabulation is known as treble tabulation. A manifold tabulation occurs when more than four characteristics are shown simultaneously. Following example will give a clear understanding:

Simple Table

Number of Students in ICFAI University in the Age Group	
Age group	No. of Students
15-25	….
25-35	….
35-45	….
45 –55	….
Total …..	

Two Way Table

Number of Students in ICFAI in Different age Group According to Sex			
Age group	Students		Total
	Female	Male	
15-25	…	…	…
25-35	…	…	….
35-45	…	…	….
45-55	…	….	…
Total	….	….	…..

Higher order or Manifold tables are prepared when three or more characteristics are to be represented in the same table. It gives information on large number of inter-related characteristics of a given phenomenon simultaneously. In such table it is necessary to decide the order of precedence among the characters to be classified with respect to their relative importance. But as the character to be shown in the table increases, the table becomes more and more confusing. So in order to avoid confusion, a table should represent only the four characters. When the character to be represented is more than four, we should construct another table for showing the relationship between them. The example of manifold table is given below:

Manifold Table

Number of Student in ICFAI University according to Age, Sex & Programs							
Age group	Students						Total
	BBA		MBA		CFA		
	F	M	F	M	F	M	
15-25	-	-	-	-	-	-	-
25-35	-	-	-	-	-	-	-
35-45	-	-	-	-	-	-	-
45-55	-	-	-	-	-	-	-
Total	-	-	-	-	-	-	-

General and Special Purpose Table

General purpose table is a convenient way of presenting systematically and chronologically arranged data or the information in a suitable form for general use or reference. It is also known as reference table or informative tables or repository tables. They contain detailed information. Example of such tables are the reports published by government agencies, pay rolls of industrial houses, etc.

Special purpose table provide information of analytical nature and for particular discussion. It is prepared for making comparative studies and studying the relationship and importance of figures provided by the data. It is also known as summary tables. As they are derived from the general table, they are also known as derivative table. In order to facilitate comparison, the table provides interpretative figure such as ratios, percentages etc.

Distinction between Classification and Tabulation

Classification and Tabulation go together and are not distinct from each other. Classification is the primary step in the process of Tabulation. It means that various items with common characteristics are brought together and are displayed in a table. This makes reader to understand their relationship. Thus, Tabulation includes Classification. In other words, the scope of Tabulation is much wider to Classification.

TABLE REVIEW

An experienced person must review a table in its form, content, validity and clerical accuracy as it is difficult for the main person to make a through analysis of all the aspects of the table. The reviewer can make an analysis on following aspects:

i. Title clarity

ii. Pertinence of all the entries

iii. Uniform subject matter

iv. Proper arrangement of classification to facilitate comparison

v. Emphasis of important points

vi. Source of the data

vii. Notations about peculiarities of the data etc.

FREQUENCY DISTRIBUTIONS

Before actually studying the concept of frequency distribution, let us study the meaning of the term 'Variable'.

A variable is a characteristic, which vary in magnitude in a frequency distribution. For example, the marks of the students, height or weights of the students, wages of workers, profits of the company, etc. The variables are of two types. (i) Continuous Variable, and (ii) Discrete variable.

A continuous variable is a variable, which is capable of manifesting all possible values, i.e., integral and fractional within the given specified range such as height and weights of the students, age of students. Thus, a continuous variable is capable of passing from any given value to the next value with infinitely small gradations.

A discrete variable is a variable that is exact and finite and cannot be expressed in fractions or decimals. For example: number of rooms in a house, number of children in a family etc.

The number of times a particular variable gets repeated is called frequency. A frequency distribution is constructed on the basis of Statistical Series. A statistical series is a succession of quantitative values. It refers to some logical arrangement of data based on size or magnitudes or characteristics, as the case may be.

A statistical series is divided into:

- **Individual Series** – it is a raw data i.e., the data without frequency. It can be seen from the following example:

10	0	20	30	40	50	60	70	20
30	40	0	20	50	80	40	50	90
60	30	20	70	60	50	30	30	70
80	90	70	90	40	50	40	30	

- **Discrete Series** – it is a series that deals with discrete variables.
- **Continuous Series** – it is a series that deals with continuous variable.

A frequency distribution means a series where a number of observations with similar or closely related values are put in separate groups or classes each of them being in the order of magnitudes.

According to Croxten and Cowdon, "a frequency distribution is a statistical table which shows the set of all distinct values of the variable arranged in order of magnitude, either individually or in groups, with their corresponding frequencies side by side".

It is of two types. They are:

i. **Discrete Frequency Distribution** – In this form of distribution, the frequency refers to a finite variable such as number of rooms in a house, number of students in a class etc.

ii. **Continuous Frequency Distribution** – In this form of distribution, the frequencies refer to groups of values. It is necessary for such variables which takes the fractional form or where the exact measurement is not possible. For example: weekly wages, height etc. The details regarding their formation are discussed below:

Formation of Discrete Frequency Distribution

The process of preparing a discrete frequency distribution is very easy and simple. We have to count how many times a particular value of the variables is repeating in the given data which is called as the frequency of that class. To facilitate the counting, we prepare a column known as 'Tally bars' or tally marks or tallies.

In the first column, we place all the possible values of the variable from lowest to highest. In the second column, put the vertical lines called as Tally Bars (|||) against the particular value of the variable whenever it occurs. When a particular value occurs for four times, for the fifth occurrence put a cross tally bar (/) on the first four Tally Bars (||||) and count this as five. The block of five is created for facilitating counting and for providing space in between each block. After putting the tally bars for all the values, we count the number of bars and write it in the third column against the corresponding number, thus we get frequency. The representation of the data in this manner is known as *Discrete or ungrouped frequency distribution*. It will be clearer with the following illustration:

Illustration 1

Prepare a frequency distribution for the example given under individual series i.e., following data relating to 35 students and the marks secured by them.

10	0	20	30	40	50	60	70	20
30	40	0	20	50	80	40	50	90
60	30	20	70	60	50	30	30	70
80	90	70	90	40	50	40	30	

Solution

Discrete Frequency Distribution

Marks	Tally Bars	Frequency
0	||	2
10	|	1
20	||||	4
30	|||| |	6
40	||||	5
50	||||	5
60	|||	3
70	||||	4
80	||	2
90	|||	3
	Total	N = 35

Formation of Continuous Frequency Distribution

When the identity of the unit in respect of which, the information is collected is not relevant nor is the order in which the observation occurs, then the first step of condensation involves classification of data into different classes and then recording the number of observation in each class. The presentation of data into continuous classes with their corresponding frequencies is known as continuous frequency distribution. The following technical terms should be borne in mind when the data is classified according to continuous frequency distribution.

CLASS LIMIT

The lowest and the highest value included in the class are called as class limits. For example if the class is 10-20 then 10 is the lowest value and 20 is the highest value. 10 is the lower boundary of the class and is called as lower limit. The lower limit is the value below which there is no item in the class. 20 is the upper boundary of the class and known as upper limit. The upper limit is the value above, which no items belong to that class.

CLASS INTERVAL

The class interval of a class is the difference between the lower limit and the upper limit of that class. For example, in a class of 5-15, the class interval is 10 (15-5). The width of the class is an important aspect to be considered while constructing a frequency distribution. It depends on number of factors such as range of the data, the number of classes needed to be formed, etc. A class interval can be estimated by following formula:

$$i = \frac{L - S}{k}$$

Where,

k – The number of classes

L – Value of largest item

S – Value of smallest item.

For example, if the salary of 50 employees in a factory varies between 500 to Rs.2,500 and desire to have 5 classes, then the class interval would be,

$$i = \frac{L - S}{k}$$

Where,

L = 2,500; S = 500 and k = 5

$$i = \frac{2,500 - 500}{5} = 400$$

Therefore, the starting class will be 400-800, 800-1200 and so on.

The number of classes 'k' can be fixed arbitrarily depending on the problem under study or can be decided through an approximate formula given by Prof. Sturges known as Sturger' Rule. According to him, the number of classes can be determined with following formula:

$$k = 1 + 3.322 \log N$$

Where, N = the total number of observation, log = logarithms of the number and the value obtained is rounded off to the next higher integer. If the total number of observation is 10 then the class interval will be,

$$k = 1 + 3.322 \log 10 = 4.322 \cong 4$$

For determining the magnitude of the class interval Sturges suggested following formula:

$$i = \frac{\text{Range}}{1 + 3.322 \log N}$$

Where, Range is nothing but the difference between the largest and smallest value.

Types of Class Interval

Data can be classified according to class interval in two ways i.e.,

i. Exclusive Method

ii. Inclusive Method.

- **Exclusive Method:** When the class intervals are such that the upper limit of one class becomes the lower limit of another class, such classification is known as exclusive method of classification. In other words, when the upper limit of a class is excluded from the respective class and is included in the immediate next class, such classes are known as exclusive classes. The following data is an example of exclusive method:

Weekly Wages (in Rs.)	No. of Workers
0-20	41
20-40	51
40-60	64
60-80	38
80-100	6

0-20 i.e. $0 \le X < 20$

20-40 i.e. $20 \le X < 40$

This method ensures continuity of the data, as the upper limit of the class becomes the lower limit of immediate next class. This method is popularly used in practice but is confusing for a layman without the knowledge of statistics.

- **Inclusive Method:** When both upper and lower limit of the class are included in that class itself, then the method of classification is known as Inclusive method. The following distribution is an example of inclusive method:

Weekly Wages (in Rs.)	No. of Workers
0-19	41
20-39	51
40-59	64
60-79	38
80-99	6

$0\text{-}19 \Rightarrow 0 \le X \le 19$

$20\text{-}29 \Rightarrow 20 \le X \le 29$

In this example, we include only those workers whose wages are in between 0 and 19. Suppose if the wages of a worker is Rs.20 then it will be included in the next class.

The use of inclusive or exclusive method depends on whether the variable under distribution is a discrete variable or a continuous variable. Generally, inclusive method is used for discrete variables and exclusive method is used for continuous variables.

CLASS FREQUENCY

Class frequency is nothing but the number of observations corresponding to a particular class. It is also known as frequency of the class. In the following illustration the class frequency 40-60 is 64 which implies that 64 workers are receiving wages between Rs.40 and Rs.60 per week. If the individual frequency of all the classes is added we get the total frequency i.e. 200. It implies that we have studied wages of 200 workers.

Weekly Wages (in Rs.)	No. of Workers
0-20	41
20-40	51
40-60	64
60-80	38
80-100	6

MID-POINT OF A CLASS

Mid point is the value which is half-way between the upper class and lower class limit of a class interval. It is calculated by following formula:

$$\text{Mid-point} = \frac{\text{Upper limit of a class + Lower limit of a class}}{2}$$

CONVERSION INTO CONTINUOUS SERIES WHEN MID POINTS ARE GIVEN

In this case the difference between the two mid-point is taken and is divided by 2. The resultant value after such division is deducted from the mid points to find the lower limit and added to the mid-point to find the upper limit of a particular class. This can be seen with the help of following example:

Example

Mid-point (m)	No. of Companies (f)
115	12
125	20
135	25
145	28
155	15
	100

The common difference between the mid-points is 10 (125 – 115 =10 or 135 – 125 = 10 etc.). To get the lower and upper limits of the respective classes the following procedure is adopted.

Mid-point (m)	No. of Companies (f)
(115-10) – (115+10)	12
(125-10) – (125+10)	20
(135-10) – (135+10)	25
(145-10) – (145+10)	28
(155-10) – (155+10)	15
	100

Mid-point (m)	No. of Companies (f)
105 – 125	12
115 – 135	20
125 – 145	25
135 – 155	28
145 – 165	15
	100

FACTORS TO BE CONSIDERED WHILE CONSTRUCTING A CONTINUOUS FREQUENCY DISTRIBUTION

The construction of frequency distribution depends on the nature of the data and the classification objective. However, in order to ensure meaningful classification of data following consideration should be borne:

- The number of classes should be between 5 and 20 preferably. However, there is no hard and fast rule but it should not be less than 5, otherwise the classification may not reveal the essential characteristic.
- The class interval of the class should be preferably either 5 or in multiples of 5. The values of class intervals such as 3, 7, 11, etc., should be avoided.
- The lower limit of first class should be either 0 or 5 or multiple of 5. For example, if the lowest value of the data is 27, then we can have a class interval of 5 or 10 so that the first class will be either 25-30 or 20-30.
- Exclusive method should be adopted to ensure continuity and obtain correct class interval. But when inclusive method is adopted, then adjustments should be made for getting correct class interval and continuity. The correction factor or adjustment is calculated by following formula:

$$\text{Correction factor} = \frac{\text{Lower limit of 2nd class} - \text{Upper limit of 1st class}}{2}$$

The value obtained by above formula is added to the value of upper limit and deducted from the value of lower limit of the class. The following illustration will clear it:

Illustration 2

From the following data you are required to draw a frequency distribution based on exclusive method:

Weekly Wages (in Rs.)	No. of Workers
0-19	41
20-39	51
40-59	64
60-79	38
80-99	6

Solution

For ensuring continuity we take the difference between the upper limit of the first class (19) and the lower limit of 2nd class (20) which is 1. Divide it by 2 so we get 0.50 which is the Correction factor. Now deduct 0.50 from lower limit of all classes and add 0.50 to the upper limit of all the classes. Now the adjusted class will be:

Weekly Wages (in Rs.)	No. of Workers
0.5-19.5	41
19.5-39.5	51
39.5-59.5	64
59.5-79.5	38
79.5-99.5	6

RELATIVE FREQUENCY DISTRIBUTION

A frequency distribution may be converted into relative frequency distribution in order to know the percentage of the total number of observations in each class. The total value of the relative frequency is always equal to 1. The frequency of each class is divided by the total frequency in order to get relative frequency. Lets take an example to have a clear understanding:

Marks	No. of Students	Relative Frequency
0-10	2	0.1(2 / 20)
10-20	5	0.25
20-30	4	0.20
30-40	5	0.25
40-50	4	0.20
Total	20	1.00

BIVARIATE OR TWO-WAY FREQUENCY DISTRIBUTION

Frequency distribution involving only one variable is known as Univariate frequency distribution. Till now our study was confined to such distribution only. But sometimes it is necessary to study two variables for the same population simultaneously, such as data relating to the marks in accountancy and marks in economics or the age of the students and their heights or weights, etc. When the data is classified on the basis of two variables, such a frequency distribution is known as Bivariate frequency distribution. The consideration remains same in bivariate distribution. The data relating to one variable say X is grouped into *m* classes and the data relating to other variable say Y is grouped into *n* classes. Then the bivariate table consist of *m x n* cells. The bivariate table is also known as correlation table. For obtaining the bivariate frequency table, we use the different pairs of the values (x,y) of the variables and use tally bars to find the frequency of each cell. The format of a bivariate frequency table is given below:

		Class Interval of X-series	Frequency
		x_1 x_2 … x_n	
Class-Intervals of Y-series	y_1 y_2 ⋮ y_n	f(x,y)	f_y
Total Frequency of X		f_x	Total – N $\sum f_x = \sum f_y$

SUMMARY

- The first step towards further processing of data is classification. Classification is grouping of related facts into classes. It is nothing but arrangement of data on the basis of nature, objectives and scope of enquiry.
- The data is classified on geographical. Chronological, qualitative and quantitative basis tabulation. It simplifies presentation facilitates comparisons and statistical analysis.
- A statistical table is the systematic arrangement of data in rows and column's according to some salient features.

Chapter V

Diagrammatic and Graphic Presentation

After reading this chapter, you will be conversant with:

- Significance of Diagrams and Graphs
- Diagrams
- Types of Diagrams
- Graphs
- Types of Graphs
- Techniques of Constructing Graphs
- Difference between Diagrams and Graphs
- Limitations of Diagrams and Graphs

Introduction

Earlier we have learnt the techniques or devices for summarizing and presenting the statistical data in a systematic and readily comprehensible manner. But however, such a presentation may not be interesting for a common layman, because it is confusing and does not convey the message for which it is meant. The other convincing, appealing and easily understood method in which the statistical data can be presented is diagrams and graphs. The statistical data can be displayed pictorially in various ways such as points, lines, graphs and maps etc. Selection of the best method out of several methods is a difficult task, because it requires artistic talent and imagination while preparing the diagrams and graphs. The present chapter deals with all the aspects of diagrammatic and graphical presentation of the data.

SIGNIFICANCE OF DIAGRAMS AND GRAPHS

The diagrams and graphs are extremely useful and have the following advantages:

- Diagrams and graphs give bird-eye view of the entire given set of numerical data by presenting the data in a simple form. The information presented by diagrams and graphs is easily understood because a picture is worth 10,000 words. Thus, diagrams and graphs provide visual aid.
- When compared to numerical data, diagrams and graphs are more attractive and fascinating to eyes. Figures in the form of picture are more appreciated by eye and leave a lasting impression on the mind when compared to dry and cold statistical figures.
- A layman without any statistical background or knowledge can easily understand diagrams and graphs because it has visual appeal.
- Diagrams and Graphs are used extensively for presenting statistical data in fairs, conferences, exhibitions, board meeting etc. Also, the diagrams and graphs have universal applicability.
- They register a long lasting and meaningful impression on the mind when compared to those created by figures in tabular form. They save time and have a memorizing effect because it is grasped quickly. It also helps in drawing the inferences quickly.
- They facilitate accurate comparison of the data relating to different periods or different regions, and also help in studying the relationship between them. They exhibit the information that might be lost amid the detailed numerical tabulation.

DIAGRAMS

A diagram refers to pictorial representation of data in the form of bars, circles, maps etc., and does not include graphs.

Rules for Construction of Diagrams

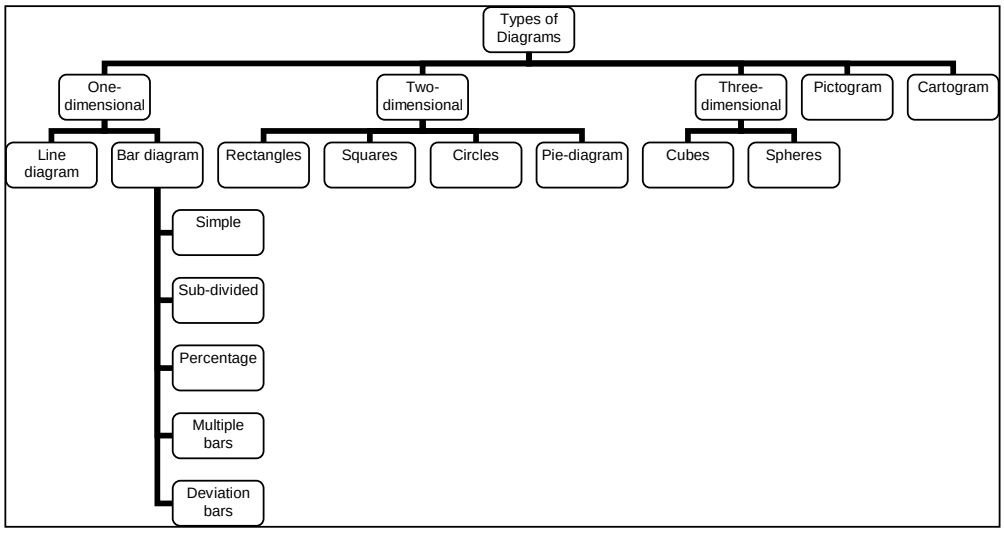
- **Neatness:** The diagram should be neat and clean. As the diagram has visual appeal, fascinates eyes, and leaves a memorizing impression on the mind, therefore it should be drawn neatly with lettering and proper size. It can be made more attractive by using different colors, shades, dotted lines, etc.
- **Title:** Like tables every diagram must have a suitable title for indicating the subject matter and various facts. The title of the diagram should convey the main idea in few words. In other words, it should be self-explanatory, clear and unambiguous. However, the brevity should not be at the cost of clarity. The title may be displayed either at the top or bottom of the diagram neatly.

- **Proportion between Width and Height:** A proper proportion between the width and height (dimension) of the diagram should be maintained otherwise the diagram may give an ugly look. Lutz has suggested a standard or rule known as 'root two' in his book entitled 'Graphic Presentation'. The root two rule is 1 to $\sqrt{2}$ or ratio of 1 on the smaller side to 1.414 on the larger side. Generally, a diagram should be drawn in the middle of the page.
- **Selection of Scale:** Selection of appropriate scale is an important factor in the construction of a diagram. The scale should be selected with great caution because the same numerical data plotted on different scale may give the diagram with different size and time leading to misleading interpretations. The general rule says that the scale should have value in even number or in the multiples of five or ten. Both on the vertical and horizontal axis, the scale should be indicated and should specify the size of the unit such as million tones, rupees in crores, etc.
- **Footnote:** A footnote should be given at the bottom of the diagram for clarifying anything in the table.
- **Index:** Various types of colors, shades, lines, etc., used in the construction of a diagram should be specified in the index in order to give the reader a clear understanding of the diagrams.
- **Simplicity:** The diagram should be as simple as possible so that it is easily understood by the reader and a layman who does possess any statistical and mathematical knowledge. The diagram should not be overloaded with too much information otherwise it will be difficult to grasp and becomes confusing. As such, it is advisable to draw several simple diagrams, which will be more effective than one complex diagram.
- **Preference in Preparation:** The preparation of vertical diagrams must be given preference to the preparation of horizontal diagrams.
- **Accuracy:** The diagram should represent accuracy of the data. Accuracy should not be put at cost to the attractiveness of the diagram.

TYPES OF DIAGRAMS

In practice, a large variety of diagrammatic devices are in use for presenting the statistical data.

Figure 1



For the sake of simplicity, the diagrams are divided as follows:

i. One-dimensional diagrams such as line diagrams and bar diagrams,
ii. Two-dimensional diagrams such rectangles, squares, circles or pie-diagrams,
iii. Three-dimensional diagrams such as cubes, spheres etc.,
iv. Pictogram, and
v. Cartogram.

One-dimensional Diagrams

The one-dimensional diagram can be of two types – Line diagram, and Bar diagram.

LINE DIAGRAM

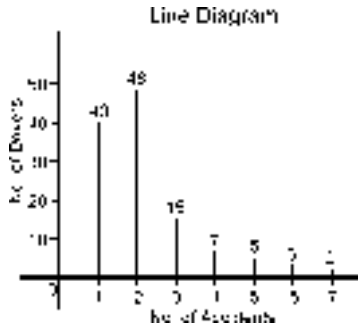
The simplest of all the diagrams is the line diagram. In this, we draw vertical lines which are equal to the frequency. The variable X is presented on the x-axis and its corresponding frequency is presented on the y-axis. These diagrams are not attractive, but facilitate comparison.

Illustration 1

The following data relates to the number of road accidents in a year. Represent it by a line diagram.

No. of Accidents:	0	1	2	3	4	5	6	7
No. of Drivers:	50	40	48	15	7	5	3	2

Solution



BAR DIAGRAM

The most commonly used device for presenting the statistical data is bar diagrams. A bar is a thick line, consists of group of equidistant rectangles whose values are represented by length or height of the rectangle and not by width. For this reason, this diagrams are also known as one-dimensional diagrams. A bar diagram is popularly used when compared to line diagram because of the following advantages:

1. They are easily understandable.
2. They are simple and easy to draw.
3. It can be effectively used for large number of items.

While drawing a bar diagram, the following points should be considered:

- All the bars should have uniform width throughout the diagram depending on the number of bars to be drawn and the space required for it.
- The gap or spacing between the different bars should be uniform to make it more attractive and elegant.
- The bars should be constructed on the same base line. It may be either vertical or horizontal.
- Figures represented by bar should be written on the top of the bar so as to enable the reader to know the precise value.

Types of Bar Diagrams

The following are the different types of bar diagrams:

i. Simple Bar diagram.

ii. Sub-divided bar diagram.

iii. Percentage Bar diagram.

iv. Multiple bar diagram.

v. Deviation bar diagram.

Simple Bar Diagram

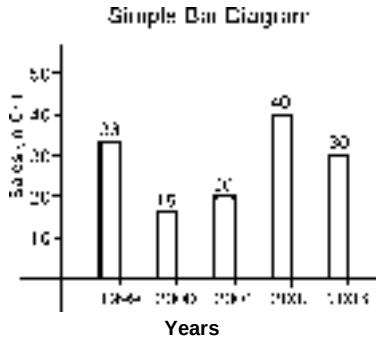
Simple bar diagram is the simplest bar diagram. It is used for representing the single variable or single classification or for comparative study of two or more values of a single variable. Simple bar diagram can be used for showing the figures of sales, profits, production, population, etc. The magnitude of the observation is represented by the height of rectangle. This facilitates the study of relationship. Simple bar diagrams may be vertical or horizontal, but vertical bars are more popularly used in practice. The only drawback of this diagram is that it represents only one classification.

Illustration 2

The following data gives the sales of a company over the period of 5 years:

Years:	2003	2004	2005	2006	2007
Sales (‘Crore):	33	16	20	40	30

Solution



Sub-divided Bar Diagram

The serious drawback of simple bar diagram is that it represents only one variable such as the total number of students in a class during the 3-year undergraduate course, but it fails to show the faculty-wise or sex-wise distribution of the students. In such a case, we use sub-divided bar diagrams. In sub-divided bar diagram, each bar representing a given phenomenon is further divided into parts, components or classes. The sub-divided part occupies a part proportional to its share in the total bar.

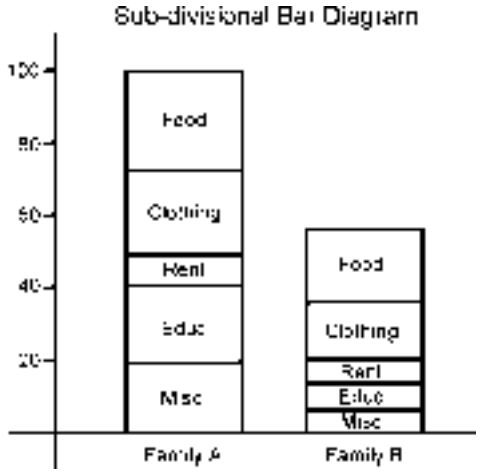
In this type, we first draw a bar diagram for the total and then divide it into various components in proportion to its share. Different bars are represented in different colors or shades or crossing or dotting and an index is given for explaining the difference along with the diagram. The component can be represented either in absolute form or in the percentage form.

Illustration 3

The following data relates to the expenditure of two families X and Y whose income is Rs.1,000 and Rs.500 respectively. Represent the data in a suitable form:

Expenditure	A	B
Food	300	200
Clothing	250	120
Rent	50	100
Education	380	40
Miscellaneous	20	40

Solution



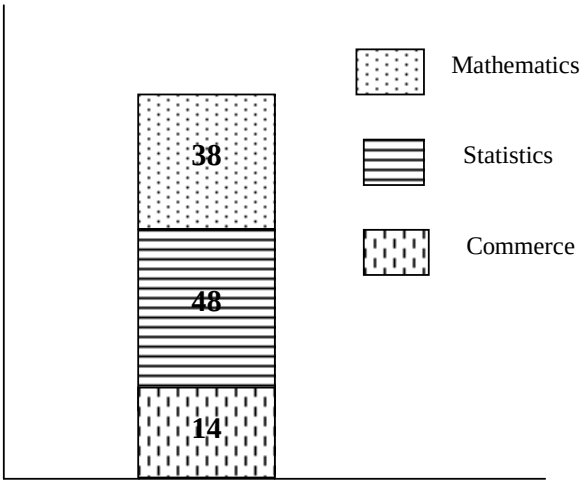
Percentage Bar Diagram

This is similar to sub-divided bar diagram. Even in this diagram, the data of a particular period or particular variable presented through a single bar, but in terms of percentages. It facilitates comparison. It can be represented with the following illustration:

Illustration 4

A Student acquired 38, 48 and 14 percent of marks in Mathematics, Statistics and Commerce respectively. You are **required** to draw a percentage bar diagram based on the above data.

Solution



Multiple Bar Diagram

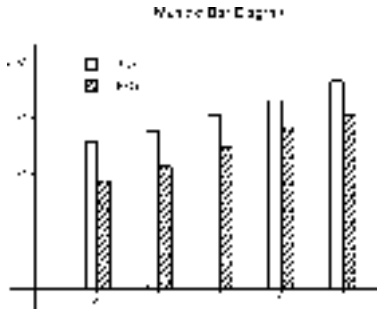
The only drawback of simple bar diagram is that it represents the single or one characteristic. In order to represent two or more inter-related variables graphically, we make use of multiple bar diagrams. It is drawn in the similar manner as we draw the simple bar diagrams. We use different colors, shades, dotting and crossings for distinguishing the different bars. An index is given along with the diagram to that effect.

Illustration 5

The following data gives sales (in thousand of rupees) of two companies – Alpha and Gamma company.

Year	Sales ('000 rupees)	
	Alpha	Gamma
1999	120	90
2000	140	100
2001	150	110
2002	160	130
2003	175	145

Solution



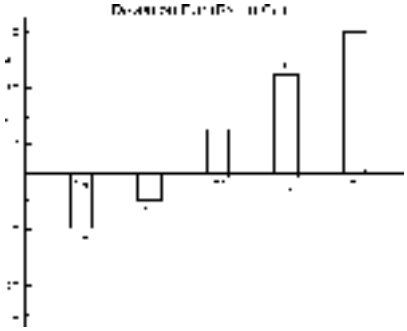
Deviation Bar Diagram

The net quantities which have positive and negative values such as surplus and deficit, profit and loss, net of import and export, can be represented graphically through deviation bars. The positive deviations are represented above the base line and the negative deviations are represented below the base line by bars. These bars are also known as Bilateral bar diagrams. It can be clearly explained by the following illustration:

Illustration 6

The following illustration shows the profit and losses made by nationalized banks during the last 5 years.

Solution



Broken Bars

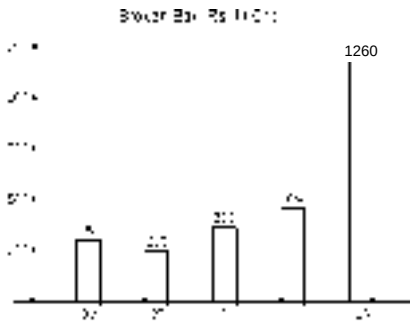
When the data is having wide variations in the values i.e., the data contains large observations along with small observations in that case, Broken Bar diagrams are used. The vertical scale makes the small bars too small and is not of much use in broken bars because it does not reveal the true characteristic of the given data. So, in order to make smaller bars attractive, we broke the larger bars at the top.

Illustration 7

The following data relates to additional taxes paid by the company A during the last 5 years.

Years:	2003	2004	2005	2006	2007
Sales (Rs. in Cr)	280	250	350	450	1260

Solution



Two-Dimensional Diagrams

Bar and line diagrams are one-dimensional diagrams in which only one dimension represents the magnitude of observation. But, in a two-dimensional diagram, the magnitude of the observation is represented by both the height and width of the bar. The two-dimensional diagrams are also called area diagram or surface diagram because they represent the magnitude of observation by its area. The two-dimensional diagrams which are commonly used are:

i. Rectangles,

ii. Squares,

iii. Circles, and

iv. Pie diagram.

Rectangles

Rectangle is a two-dimensional diagram based on the principle of area. Mathematically, the area of rectangle is determined by multiplying the length with its breadth. So, both the dimensions are considered while constructing a rectangle. There are two methods in which rectangle diagrams are represented. In the first method, we represent the figures as given and then sub-divide them into various components. The second method is percentage basis, which is used for representing the relative magnitude of two or more sets of data along with its components.

Like bars, rectangles are also placed side-by-side with equal spacing between them. Rectangle diagram is a modified version of bar diagram and gives more detailed information when compared to bar diagram.

Illustration 8

The following data relates to the production of product A and B in a factory. Represent the data in a rectangular diagram form

Details	Product A	Product B
Units produced	1000	1500
Cost of raw material	Rs.5000	Rs.6000
Other production expenses	Rs.3000	Rs.4000
Profit	Rs.1000	Rs.2000

Solution

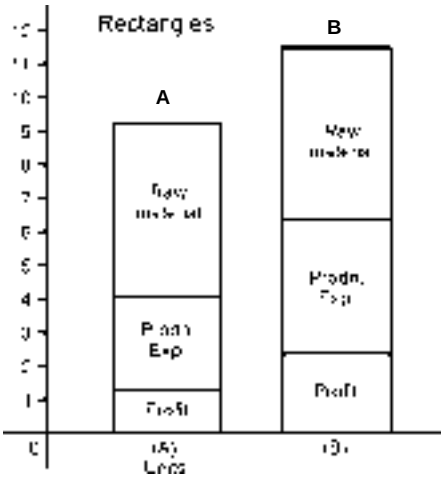


Illustration 9

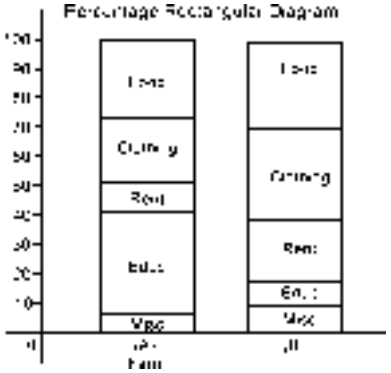
The following data relates to the expenditure of two families X and Y whose income is Rs.1,000 and Rs.500 respectively. Represent the data in a percentage diagram.

Expenditure	A	B
Food	300	200
Clothing	250	120
Rent	50	100
Education	380	40
Miscellaneous	20	40

Solution

First we convert the above data into percentage.

Expenditure	Family A			Family B		
	Rs.	%	Cumulative	Rs.	%	Cumulative
Food	300	30	30	200	40	40
Clothing	250	25	55	120	24	64
Rent	50	5	60	100	20	84
Education	380	38	98	40	8	92
Miscellaneous	20	2	100	40	8	100
Total	1000			500		



Square

When the values of items vary widely, it is difficult to use rectangular method. So, the other method, which can be used for comparing the values that differs widely from one another is a square diagram method, for example, the population of India at different times, or population of different countries at a particular time, etc. The square diagram is a two-dimensional diagram, which represents the given value by the area of the square. The area of square is given as side x side or square of the side. So, the side should be proportional to the square root of given observation.

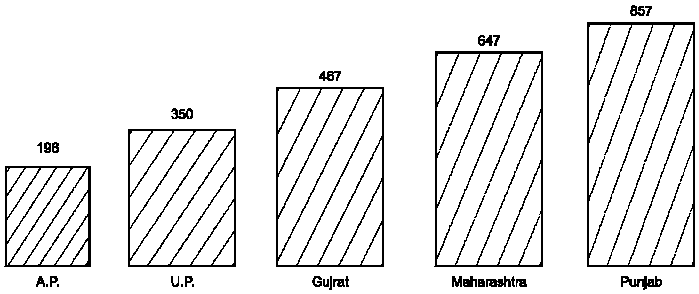
The first step in the construction of a square diagram is to obtain the square root of the observation and then draw the square with sides proportional to square root.

Illustration 10

Draw a square diagram for the following data relating to the yield per hectare in different states of India:

States	AP	UP	Gujarat	Maharashtra	Punjab
Yield per hectare	196	350	467	647	857

Solution



Circle

The other method of preparing a two-dimensional diagram is circle. It is alternative to squares that are used for same purpose. The total and its components can be shown in this diagram. The area of circle is proportional to the square of its radius i.e. πr^2 where π = 22/7 and r is the radius of the circle. So, the length is taken as the radii of the circle. While constructing the circle, we obtain the square root of figures and then radii is calculated by dividing the pie value and take the square root.

Pie Diagram

The most popularly used diagrams in practice are the pie diagrams for showing the percentage breakdown. For example, pie diagram can be used for showing the distribution of the government expenditure into various heads such as industry, agriculture, defense, transport, telecommunication, etc. Similarly through pie diagram, we can show the allocation of family expenditure spent on food, clothing, rent, education, saving, etc. The diagrams are called pie diagrams because the components resembles to slices cut from the pie. The pie diagrams should be used on percentage basis for comparison. While constructing a pie diagram, we follow logical sequence i.e., the largest component is started at 12 O'clock position, followed by other components in the descending order.

The following steps are followed while constructing a pie diagram:

- Convert the various component value into corresponding degrees on circle.
- Draw circles of appropriate size and radius depending on the available space and other factors.
- Now draw any radius and take it as a base line for drawing angles at the center that is equal to the degree of first component. Draw a new line at the center to form this angle which will touch the circumference. This sector represents the proportion of the first component.
- We can use different shades, colors, dotting, etc., for representing the various components and then give descriptive labels either inside or outside of the circle.

Illustration 11

Draw a pie diagram for the following data of a five-year plan for public sector outlays:

Agriculture and Rural Development	12.9%
Irrigation	12.5%
Energy	27.2%
Industry	15.4%
Transport	15.9%
Others	16.1%

Solution

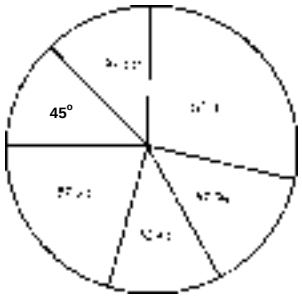
The angle at the center is calculated by the following formula:

$$\frac{\text{Percentage Outlay}}{100} \times 360 = \text{Percentage Outlay x } 3.6^0$$

Computation for Pie Diagram

Sectors	Percentage (%)	Angles
Agriculture and Rural Development	6.45	12.9 x 3.6 = 46.44^0
Irrigation	6.25	12.5 x 3.6 = 45^0
Energy	13.6	27.2 x 3.6 = 97.92^0
Industry	7.7	15.4 x 3.6 = 55.44^0
Transport	7.95	15.9 x 3.6 = 57.24^0
Others	8.05	16.1 x 3.6 = 57.96^0

Now draw a circle and divide it into 6 parts according to the degree of angle.



Three-dimensional Diagrams

Three-dimensional diagrams are those diagrams in which the three dimensions – height, length and breadth are taken into account. These diagrams are also known as volume diagrams because its volume represents the given magnitude. It can be presented in the form of cubes, cylinders, spheres etc. These diagrams are used when the range of difference between the largest and smallest magnitude is very large.

Limitations: The limitations of three-dimensional diagrams are similar to two-dimensional diagrams. These diagrams are difficult to read and hence are not used much in presenting the statistical data.

Pictograms

The statistical data can also be presented through a technique known as Pictogram. It is popularly used for presenting the statistical data and facts to the layman who does not possess any statistical and mathematical background. The pictograms are represented through a pictorial symbol, which should be carefully

selected. A picture is attractive and appealing to the eyes of the readers and has the lasting and memorizing impression on his mind. They are widely used by the government and social agencies for presentation of the data to the general public.

Merits

Pictograms are more attractive when compared to other type of diagrams. They attract the attention of masses and increase their interest in the information being represented. But, the construction of pictograms is difficult. This method is also not free from limitation. Still, it gives the overall picture without minor details etc.

Cartograms

When the statistical data is represented through maps along with various diagrams, the technique is known as Cartograms. They depict the quantitative facts on geographical basis such as rainfall in different regions, population of different countries, etc. Dots, different colors or shades represent the quantitative. Cartograms are easy and simple to understand.

Choice of a Diagram

We studied different types of diagrams that are used for presenting the statistical data with their relative merits and limitations. One particular diagram will not suit all the situations. Therefore, choice has to be made, which requires skills intelligence and expertise. The choice basically depends on the nature of the data and object of the presentation. The diagram for presenting the data should be chosen with utmost care and diligence. Otherwise, a wrong selection of diagram will distort the basic characteristic of the data and may lead to misleading interpretations.

GRAPHS

Though diagrams are useful for visual presentation of categorical and geographical data, yet the data relating to time series and frequency distributions can be best represented through graphs. Further, graphs are more obvious, precise and accurate than diagrams and can be effectively used for further statistical analysis. A graphical mode of presentation refers to the joining of various points through lines. Graph is the tool used for this purpose.

Rules for the Construction of Graphs

i. Graph must have a clear **title** to represent the facts of the respective data.

ii. The **structural framework** of a graph should be such that the independent variables of the data should be shown, on the X-axis and dependent variables should be shown on the Y-axis.

iii. The **selection of a scale** must be done keeping in mind the nature of data collected.

iv. The **false base line** may be used to mark the fluctuations in a variable, which is relatively small.

v. For showing proportional changes **ratio or logarithmic scale** should be used.

vi. If more than one line is plotted on the graph, they should be represented by different **line designs**.

vii. The **caption** for the X-axis is placed at its centre and the **caption** for the Y-axis is placed at its top.

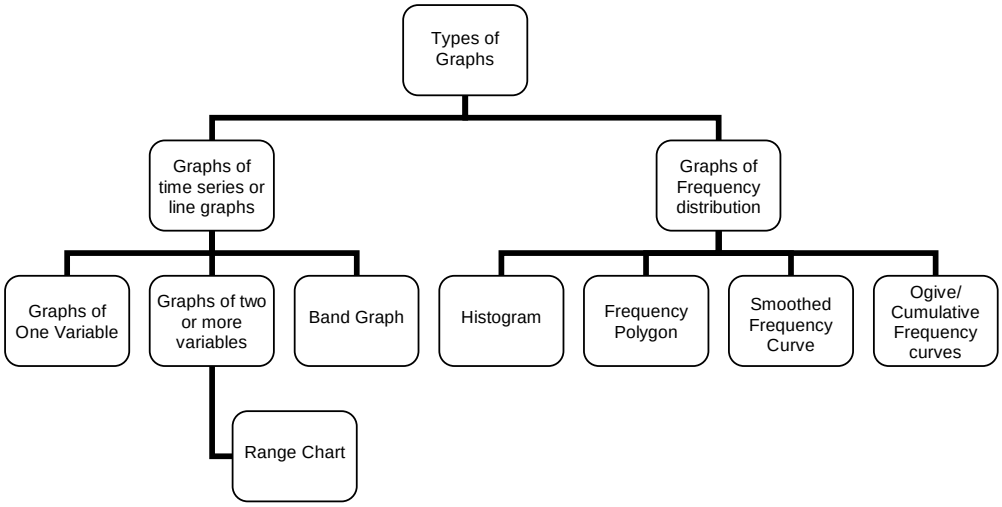
viii. An **index** should be given to show the scale and meaning of the curve.

ix. The **source of the data** collected must be indicated.

TYPES OF GRAPHS

The different types of graphs can be observed from the following figure:

Figure 2



The various graphs can be divided under the following two heads as shown in the figure 2:

- Graphs of time series or line graphs, and
- Graphs of frequency distribution.

Graphs of Time Series or Line Graphs

The series formed by values of a variable at different points of time, is known as time series. Generally, we take time on the X-axis and value of the variables on the Y-axis and join the various points by straight lines. The graph so formed is known as line graph. Many variables can be shown on the same graph and a comparison can be made. Such graphs require the least technical skill, simplest to understand, easiest to make and most adaptable to many uses.

Graphs of time series can be constructed either on a natural scale or on a ratio scale. In natural or arithmetic scale, absolute changes from one period to another are shown whereas in a ratio scale the rates of changes or the relative changes are shown.

RULES FOR CONSTRUCTING THE LINE GRAPHS ON NATURAL SCALE

- Take the time on the X-axis (horizontal) and the variables on the Y-axis (vertical).
- Begin Y-axis with zero and select a suitable scale so that the entire data is accommodated in the space available. On the arithmetic scale, equal distances must represent equal magnitude. This requirement is true for both the X-axis as well as the Y-axis separately.
- Corresponding to the time factor plot the value of the variable and join the various points by straight lines.
- If on one graph more than one variable is shown, they should be distinguished by the use of thick, thin, dotted lines, etc. Every graph should be given a suitable title. The unit of time in which the variable under consideration is measured should be clearly stated in the title.

GRAPHS OF ONE VARIABLE

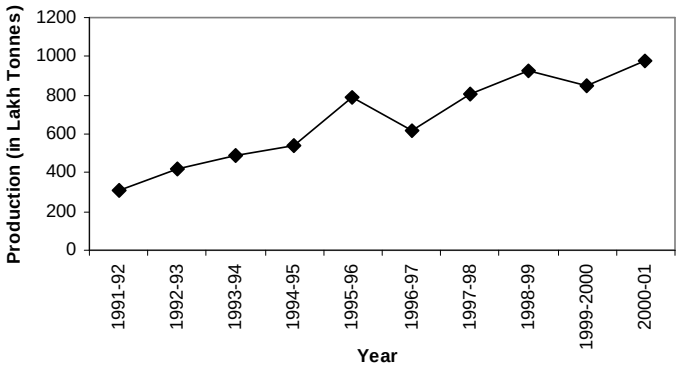
When only one variable is to be represented, Plot a graph by taking the time on the X-axis and value of the variable on the Y-axis and joining the points with straight lines. The fluctuation of this line indicates the variations in the variable, and the distance of the plotting from the base line of the graph indicates the magnitude.

Illustration 12

Represent the following data of production of a company graphically:

Year	Production in Lakh Tonnes
1991-92	305
1992-93	420
1993-94	485
1994-95	538
1995-96	785
1996-97	618
1997-98	810
1998-99	922
1999-2000	848
2000-01	981

Solution



GRAPHS OF TWO OR MORE VARIABLES

When the unit of measurement is the same, two or more variables can be represented on the same graph. But, when the number of variables is very large and they are all shown on the same graph, the chart becomes quite confusing because different lines may cut each other and make it difficult to understand the behavior of the variables. Therefore, for the sake of clarity more than 5 or 6 variables should not be represented on the same graph. When two or more variables are shown on the same graph, thick, thin, broken, dotted lines can be used to distinguish between the various variables.

Range Chart

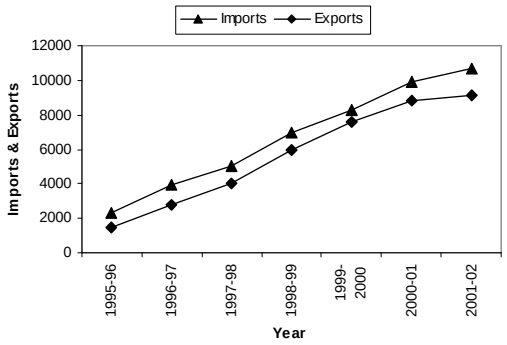
Range chart is a method of presenting the data in graphical form, which indicates the range of variation in the minimum and maximum values of a variable. This method of graphical presentation is appropriate where we have to show the minimum price of a commodity for different periods of time or the minimum and maximum temperatures, or the minimum and maximum price of shares of some companies for different periods.

Illustration 13

Show the following figures by a suitable diagram:

Year	Imports	Exports
1995-96	2,312	1,971
1996-97	3,972	2,793
1997-98	5,014	4,015
1998-99	6,995	5,923
1999-2000	8,297	7,568
2000-01	9,876	8,843
2001-02	10,681	9,982

Solution



BAND GRAPH

Band graph is a type of line graph, which shows the total for successive time periods broken up into sub-totals for each of the component parts of the total. It represents how and in what proportion the individual items comprising the aggregate are distributed. The various components parts are plotted one over the other and the gaps between the successive lines are filled by the different shades and colors. The band graph can also be used where the data is put to percentage form; the whole chart will depict 100 percent and the bands, the percentage that each component bears to the whole.

GRAPHS OF FREQUENCY DISTRIBUTIONS

A frequency distribution can be presented graphically in any of the following ways:

- Histogram,
- Frequency polygon,
- Smoothed frequency curve, and
- 'Ogives' or cumulative frequency curves.

Histogram

Histogram is one of the most popular and commonly used devices for charting continuous frequency distribution. It is a set of vertical bars whose areas are proportional to the frequencies represented. While constructing a histogram, the variable is always taken on the X-axis and the frequencies depending on it on the Y-axis. Each class is then represented by a distance on the scale that is proportional to its class-interval. The distance for each rectangle on the X-axis shall remain the same in case the class intervals are uniform throughout. The

Y-axis represents the frequencies of each class, which constitute the height of its rectangle. Histogram is widely used for graphical presentation of a frequency distribution. Histograms cannot be constructed for frequency distributions with open-end classes unless we assume that the magnitude of the first open class is same as that of the succeeding class and the magnitude of the last open class is same as that of the preceding class. It differs from bar diagrams. A bar diagram is one-dimensional, only the length is taken into consideration, but in a histogram both the length as well as the width are important; so, it is two-dimensional.

When class intervals are equal, take frequency on the Y-axis, the variable on the X-axis and construct adjacent rectangles. In such a case, the height of the rectangles will be proportional to the frequencies.

When the class-intervals are unequal, a correction for unequal class intervals must be made. The correction consists of finding for each class the frequency density or the relative frequency density.

$$\text{Frequency Density of a Class} = \frac{\text{Frequency of the Class}}{\text{Magnitude of the Class}}$$

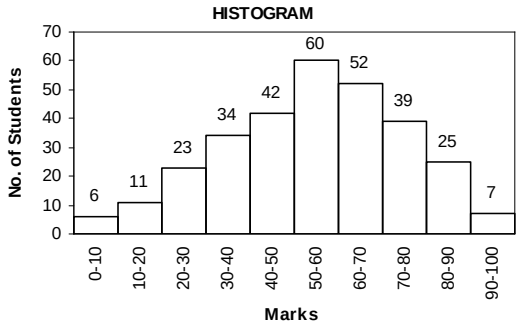
For making the adjustment, we take that class which has the lowest class-interval and adjust the frequencies of other classes in the following manner. If one class-interval is twice as wide as the one having the lowest class-interval, we divide the height of its rectangle by two, if it is three times more we divide the height the height of its rectangle by three, etc., i.e., the heights will be proportional to the ratio of the frequencies of the width of the class.

Illustration 14

Represent the following data by a histogram:

Marks	No. of Students
0-10	6
10-20	11
20-30	23
30-40	34
40-50	42
50-60	60
60-70	52
70-80	39
80-90	25
90-100	7

Solution



Frequency Polygon

A frequency polygon is a graph of frequency distribution. It has more than four sides and effective in comparing two or more frequency distributions. The frequency polygon may be constructed in two ways such as:

a. A histogram of the given data may be drawn by joining the mid-points of the upper horizontal side of each rectangle by straight lines. The figure so formed is called frequency polygon. It is an accepted practice to close the polygon at both ends of the distribution by extending them to the base line. By doing this, two hypothetical classes at each end would have to be included – each with a frequency of zero. This extension is made with the object of making the area under polygon equal to the area under the corresponding histogram.

b. The second method of constructing frequency polygon is to take the mid-points of the various class-intervals and then plot the frequency corresponding to each point, and to join all these points by straight lines. The figure obtained would exactly be the same as method one, but the only difference is that here we do not have to construct a histogram.

Frequency polygon has the following advantages over the histogram:

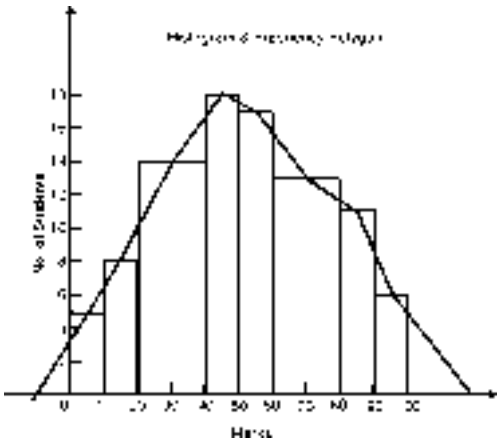
- Histogram is a two-dimensional figure, viz., a collection of adjacent rectangles whereas frequency polygon is a line graph.
- Frequency polygon is simpler than histogram.
- Frequency polygon sketches an outline of the data pattern more clearly.
- The polygon becomes increasingly smooth and curve-like as we increase the number of classes and the number of observations.

Illustration 15

Draw a histogram and frequency polygon from the following data:

Marks	No. of Students
0-10	5
10-20	8
20-40	14
40-50	18
50-60	17
60-80	13
80-90	11
90-100	6

Solution



Smoothed Frequency Curve

Smoothed frequency curve is a smooth free hand curve drawn through the various points of the polygon. The object of drawing a smoothed frequency curve is to eliminate as far as possible the random or erratic fluctuations that might be present in the data. The area enclosed by the frequency curve is same as that of the histogram or frequency polygon. The curve should look as regular as possible and sudden turns should be avoided. In the case of the data pertaining to natural phenomenon like tossing of a coin or throwing of a dice, the smoothing can be conveniently done because such data generally gives rises to symbolical curves. However, for the data relating to social, economic or business phenomenon, smoothing cannot be done effectively as such data usually give rise to skewed curves. For drawing a smoothed frequency curve, it is necessary to first draw the polygon and then smooth it out. While smoothing a frequency curve, only frequency distributions based on samples should be smoothed and only continuous series should be smoothed.

Cumulative Frequency Curves or Ogives

Sometimes, people are interested in knowing the number of workers in an organization earning less than Rs.10,000 per month or more than Rs.10,000 per month. To know the answer it is necessary to add the frequencies. When frequencies are added they are called cumulative frequencies. These cumulative frequencies are listed in a table, which is called cumulative frequency table. The curve obtained by plotting cumulative frequencies is called a cumulative frequency curve or an Ogive.

There are two methods of construction Ogive, such as:

a. **Less than Method:** In this method, we start with the upper limits of the classes and go on adding the frequencies. When these frequencies are plotted, we get a rising curve.

b. **More than Method:** In this method, we start with the lower limits of the classes and from the frequencies we subtract the frequency of each class. When these frequencies are plotted, we get a declining curve.

Utility of Ogives

Ogive is used for the following purposes:

- To determine as well as to portray the number or proportion of cases above or below a given value.
- To compare two or more frequency distributions.
- Ogives are also drawn for determining certain values graphically such as median, quartiles, deciles, etc.

Illustration 16

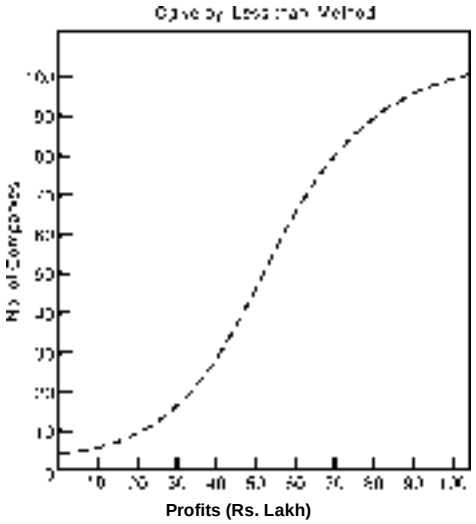
Draw less than and more than ogives from the data given below:

Profits (Rs. Lakh)	No. of Companies
0-10	4
10-20	6
20-30	9
30-40	13
40-50	16
50-60	19
60-70	17
70-80	8
80-90	5
90-100	3

Solution

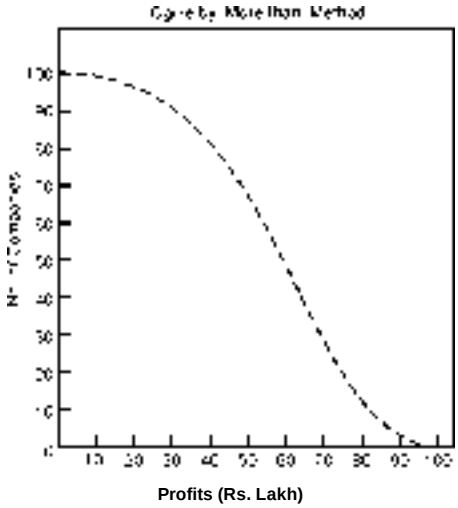
Less than ogive

Profits Less Than (Rs. Lakh)	10	20	30	40	50	60	70	80	90	100
No. of Companies	4	10	19	32	48	67	84	92	97	100



More than ogive

Profits More Than (Rs. Lakh)	0	10	20	30	40	50	60	70	80	90
No. of Companies	100	97	92	84	67	48	32	19	10	4



TECHNIQUES OF CONSTRUCTING GRAPHS

We use graph papers for constructing graphs. In the first step, two simple lines are drawn at right angle to each other; intersecting at a point, which is known as origin or zero of reference. These lines are known as coordinate axes. The horizontal line is called the X-axis or 'abscissa' and is denoted by X'0X and the vertical line is called the Y-axis or 'ordinate' and is usually denoted by Y0Y'. The whole plotting area is divided into four quadrants as shown in the figure. The distances measured towards the right or upward from the origin are positive and those measured towards the left downwards are negative. Conventionally, we take the

independent variables on the horizontal scale and the dependent variables on the vertical scale. In the case of time series, time is represented on the horizontal scale and the variable on the vertical scale. A convenient scale is chosen for each axis, which represents the unit of a variable. The choice is made in such a manner that the entire data is accommodated in the space available. The Y scale in the arithmetic line graph must begin at zero as origin. Thus, the X-axis always runs through this zero origin. In the case of natural scale equal space would represent equal amounts. However, there is no hard and fast rule about the ratio of the scale on the X-axis and the Y-axis because these would much depend upon the given data and size of the paper.

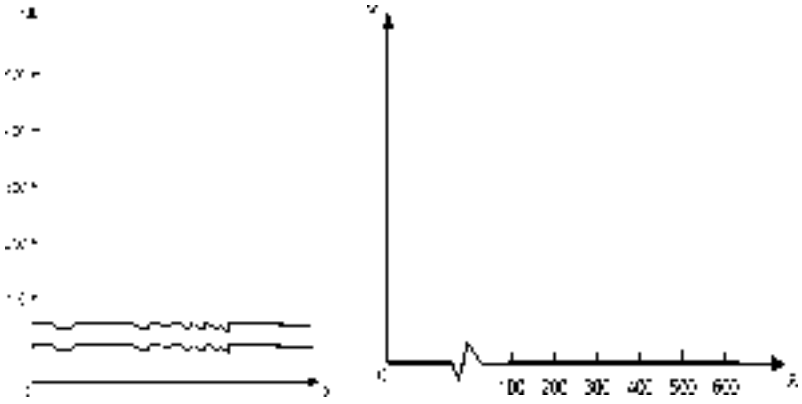
FALSE BASE LINE

The fundamental rule while constructing a graph is that the scale on the X-axis should begin from zero, where the lowest value to be plotted on the Y-scale is relatively high and a detailed scale is required to bring out the variations in all the data, starting the Y-scale with zero introduces difficulties. In this case, we have to break the Y-scale. If the zero origin is shown then drawing a horizontal wavy line or kinked or zigzag line or a vertical wavy line between zero and the first unit on the Y-scale breaks the scale. The important objectives of false base line are:

a. Variations in the data are clearly shown.

b. A large part of the graph is not wasted or space is saved by using false base.

c. The graph provides a better visual communication.



DIFFERENCE BETWEEN DIAGRAMS AND GRAPHS

There is no hard and fast rule or line of demarcation exists for distinguishing a diagram from graph. However, the following points of difference may be noted:

i. A graph is constructed on a graph paper whereas a diagram is drawn on a plain paper. A graph paper represents mathematical relationship between two variables. No such relationship is exhibited by a diagram.

ii. Various devices like bars, square, circle, cubes, etc., present a diagram, whereas in graphs, points or lines of different modes such as doted line, dashed line or combination of both are used for presenting data.

iii. Diagrams are not of much use to statisticians or research workers as they do not add anything to the meaning of the data. It provides only the approximated information and can be used for publicity and propaganda as they are attractive. On the other hand, graphs are of much use for statisticians and research workers as they are more precise and accurate than diagrams. The study of slopes, rate of change, estimation is possible because of graphs.

iv. Frequency distribution and time series are represented in graphs only. No diagrams are used for representing the time series data though it is used for representing categorical and geographical data.

v. Graph construction is much easier when compared to construction of diagrams.

LIMITATIONS OF DIAGRAMS AND GRAPHS

Diagrams and Graphs are not complete substitutes for tabular and other forms of presentations of data under all circumstances. In the words of Julin "graphic statistics has a role to play of its own; it is not the servant of numerical statistics, but it cannot pretend, on the other hand, to precede or displace the latter."

The important limitations of diagrams and graphs are:

- Only approximate values can be presented through graphs and diagrams.
- They can approximately represent only limited amount of information.
- Diagrams and Graphs are mostly intended to explain quantitative facts to the general public rather to help in analyzing the data.
- Two-dimensional diagrams and three-dimensional diagrams cannot be accurately appraised visually and, therefore, as far as possible their use should be avoided.
- Diagrams should never be accepted without a close inspection of the bonafides because they can be easily misinterpreted.

SUMMARY

- Statistical data can be displayed in the form of diagrams and graphs. The data presented in pictorial forms is more attractive and understandable.
- It facilitates accurate comparison of data relating to different periods on different regions.
- There are different types of diagrams used for presenting statistical data such as line diagrams, bar diagrams, rectangles, squares, circles, pie-diagrams, cubes, and spheres etc.
- There are different types of graphs used for presenting statistical data such as line graphs, range charts, band graphs, histograms, frequency polygons and ogives.
- Though, diagrams and graphs have certain similarities they differ in the presentation. It must be remembered that they are different tools of presentation of the data.
- Diagrams and Graphs have significance in statistical analysis, but they are not free from limitations.

Chapter VI

Measures of Central Tendency

After reading this chapter, you will be conversant with:

- Meaning and Objectives of Averaging
- Types of Averages
- Appropriateness of the Three Principal Averages
- Relationship among the Mathematical Averages
- Choice of a Suitable Average
- Additional Illustrations

Introduction

Once the data is collected, organized and presented, the next step is to use different statistical tools to these tabulated data for the purpose of analysis. The statistical tools normally used for the analysis are: measures of central tendency, dispersion, correlation and regression etc. Each tool has its own merits and limitations. The use of these tools depends on the nature of data and the desired output. In the present chapter, we will learn about the usage of the measures of central tendency and the other tools will be discussed in subsequent chapters.

MEANING AND OBJECTIVES OF AVERAGING

One of the most important objectives of statistical analysis is to get **one single value** that describes the characteristic of the entire mass of unwieldy data. Such a value is called the central value or an 'average' or the expected value of the variable. This single value is the point of location around, which individual values cluster and, therefore, called the measure of location. Since this single value has a tendency to be somewhere at the center and within the range of all values, it is also known as the measure of **central tendency**. The very purpose of computing an average value for a set of observations is to obtain a single value which is representative of all the items. It can be defined as follows:

"An average value is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of the data, it is also called a measure of central value."

– Croxton and Cowden

Thus, the average is of great significance as it depicts the characteristic of the whole group. Since an average represents the entire data, its value lies somewhere in between the two extremes, i.e., the largest and the smallest items. For this reason, an average is frequently referred to as a measure of central tendency.

Objectives of Averaging

The two main objectives are:

i. **To get a Single Value:** Measures of central value, by considering the mass of data in one single value, enable us to get a bird's-eye view of the entire data. Thus, one value can represent thousands, lakh and even millions of values. For example, it is impossible and is hardly of any use to remember the incomes of individual companies in a particular industry. However, if the average income is obtained by dividing the total income by the number of companies, we get one single value that represents the entire industry.

ii. **To Facilitate Comparison:** Reducing the mass of data to one single figure facilitates comparisons. Comparison can be made either at a point of time or over a period of time. For example, we can compare the average annual profits of different industries for a particular year, say, 1999-2000 and thereby conclude which industry is the best or we can compare the growth percentage in sales of a particular industry for different time periods and thereby conclude as to whether the results are improving or deteriorating. Such comparisons are of immense help in framing suitable and timely policies.

Requisites of a Good Average

An average should be: (a) vigorously defined, (b) easy to compute, (c) capable of simple interpretation, (d) dependent on all the observed values, (e) not unduly influenced by one or two extremely large or small values for large data, (f) fluctuate relatively less from one random sample to another, and (g) capable of mathematical manipulation.

TYPES OF AVERAGES

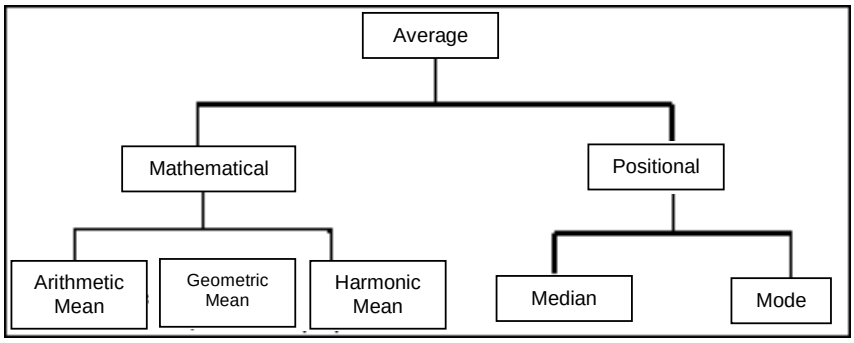
Average can be classified under mathematical and positional categories. Again, these categories are subdivided as follows:

i. Mathematical averages:
   a. Arithmetic Mean.
   b. Geometric Mean.
   c. Harmonic Mean.

ii. Positional averages:
   a. Median.
   b. Mode.

The following is the pictorial representation of various types of averages:

Figure



Though, they are segregated on the basis of their nature, the Arithmetic Mean, Median, and Mode are regarded as principal measures of central tendency/ averages. These are explained below:

Arithmetic Mean

It is called average of the given data and is obtained by attaining the total value of all the items in the data by summing their values and dividing it by number of items in the data. It is the most widely used measure to represent the entire data It is a simple average from the point of view of layman and from the point of view of statistician it is known as arithmetic mean. Since it involves mathematical aspect, it is categorized under mathematical averages.

CALCULATION OF ARITHMETIC MEAN

The calculation of arithmetic mean is done in three kinds of series i.e., individual series, discrete series, and continuous series. Let us now discuss it in detal:

Individual Series

The process of computing Arithmetic Mean in the case of individual observations is to take the sum of the values of the variable and then divide by the number of values. It is denoted by $\overline{X}$ and the formula is,

$$\overline{X} = (X_1 + X_2 + ... + X_n)/n = \left(\sum_{i=1}^{n} X_i\right)\Big/ n = \frac{\sum X}{n}$$

where,

n is the number of observations,

the variable X takes the values X_1, X_2, ... X_n, and

$\sum X$ is the sum of all the values of X.

Box 1

In statistics, the collection of all the elements under study is called a POPULATION whereas a collection of some (but not all) of the elements under study is called a sample. It is necessary to distinguish whether we are considering a population or a sample because certain formulas, like those for computing standard deviation (explained later) of a population are different from those for computing the standard deviation of a sample. Hence, population mean is denoted by,

$$\mu = \frac{\text{Sum of all the data points in the population}}{\text{Size of population}}$$

and sample mean is denoted by

$$\bar{X} = \frac{\text{Sum of all the data points in the sample}}{\text{Size of sample}}$$

Illustration 1

The following table gives the annual profits of 10 financial service companies for the year 2007-2008.

Companies	Net Profit (X) (Rs. in crore)
Ashok Leyland Finance	9.19
Classic Finance	4.27
Empire Finance	1.74
First Leasing Company	5.71
Lloyds Finance	4.80
Nagarjuna Finance	4.01
Reliance Finance	9.22
Sakti Finance	3.00
Sundaram Finance	15.16
Tata Finance	3.93
n = 10	61.03

Solution

Now, the arithmetic mean of profits of the financial services industry as represented by the above companies for the year 2007-2008 can be calculated as follows:

$$\text{Arithmetic Mean } (\bar{X}) = \frac{\Sigma X}{n} = \frac{61.03}{10} = \text{Rs.6.103 crore.}$$

This single figure of mean profit represents the profit of each financial services company under the industry on the average.

Short-cut Method: The arithmetic mean can be calculated by using an arbitrary origin. The formula for calculating arithmetic mean in this method is,

$$\bar{X} = A + \frac{\Sigma d}{n}$$

Where,

$\bar{X}$ = Arithmetic Mean.

A = Assumed mean of the data.

d = deviations from the assumed mean.

n = no. of observations.

Companies	Net Profit (X) (Rs. in crore)	(x – 8 crore)
Ashok Leyland Finance	9.19	1.19
Classic Finance	4.27	–3.73
Empire Finance	1.74	–6.26
First Leasing Company	5.71	–2.29
Lloyds Finance	4.80	–3.20
Nagarjuna Finance	4.01	–3.99
Reliance Finance	9.22	1.22
Sakti Finance	3.00	–5.00
Sundaram Finance	15.16	7.16
Tata Finance	3.93	–4.07
N = 10		–18.97

$$\bar{X} = A + \frac{\sum d}{n}$$

A = 8.00

A = 8.00, $\sum d = -18.97$, N = 10

$$\bar{X} = 8.00 - \frac{18.97}{10} = 8.00 - 1.897 = \text{Rs.6.103 crore.}$$

Discrete Series

The formula for computing mean is

$$\bar{X} = \frac{\sum fX}{\sum f} \text{ or } \frac{\sum fX}{N}$$

$$= \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{n} f_i}$$

where,

f = frequency

X = variable.

In a survey of 50 chemical industries, the following data was collected about the level of profits attained by them:

x = Level of Profit (Rs. in lakh) earned during 2007-08	f = No. of Companies that Earned X_i amount of Profit	fx
20	12	240
16	15	240
24	8	192
25	7	175
31	8	248
Total	50	1,095

The arithmetic mean is

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum fx}{N} = \frac{1,095}{50} = 21.9$$

Thus, the average profit of the chemical industry is Rs.21.90 lakh.

Short-cut Method: According to this method,

$$\overline{X} = A + \frac{\sum fd}{N}$$

X = Level of Profit (Rs. in lakh) earned during 2007-08	f = No. of Companies that earned X_i amount of Profit	(X – 20) d	fd
20	12	0	0
16	15	–4	–60
24	8	4	32
25	7	5	35
31	8	11	88
Total	50	16	95

$$\overline{X} = 20 + \frac{95}{50} = 21.9$$

Continuous Series

In continuous series, arithmetic mean can be computed by applying direct method and short-cut method:

Direct Method:

$$\overline{X} = \frac{\sum fm}{N}$$

where,

m = mid-point of class

$$= \frac{\text{Lower limit + Lower limit of next class}}{2}$$

f = frequency of each class.

N = $\sum f$ = total frequency.

A security analyst studied a hundred companies and obtained the following distribution for the dividend declared by these companies during the year 2002. Calculate the average dividend:

Dividend Declared (in percentage)	0-8	8-16	16-24	24-32	32-40
No. of Companies (f)	12	20	25	28	15

Dividend Declared	Mid-point (m)	No. of Companies (f)	fm
0-8	4	12	48
8-16	12	20	240
16-24	20	25	500
24-32	28	28	784
32-40	36	15	540
Total		100	2,112

$$\bar{X} = \frac{\sum fm}{\sum f} = \frac{2,112}{100} = 21.12\%$$

Thus, the average dividend is 21.12 percent.

Short-cut Method: In short-cut method, the formula used is:

$$\bar{X} = A + \frac{\sum fd}{N}$$

Where,

A = Assumed Mean;

d = deviations of mid-points from assumed mean; and

N = Total number of observations.

Dividend Declared	Mid-point (m)	No. of Companies (f)	d = (m – 20)	fd
0-8	4	12	–16	–192
8-16	12	20	–8	–160
16-24	20	25	0	0
24-32	28	28	8	224
32-40	36	15	16	240
Total		100		112

$$\bar{X} = A + \frac{\sum fd}{N} = 20 + \frac{112}{100} = 21.12\%$$

Correcting Incorrect Values

Sometimes, because of oversight or mistake in copying, wrong values of certain items are taken while calculating mean. To find out the correct mean, we have to deduct values of wrong items from $\sum X$ and add correct values and then divide the correct $\sum X$ by the number of observations. The result so obtained is the value of correct mean.

Formula: Corrected $\bar{X} = \frac{\text{Corrected} \sum X}{N}$

Corrected $\sum X$ = Incorrect $\sum X$ – wrong values + correct values.

Illustration 2

The mean marks of 100 students were found to be 45. Later on it was discovered that a score of 67 was misread as 76. Find out the correct mean corresponding to the correct score.

Solution

We are given N= 100, $\bar{X}$ =45

Since $\bar{X} = \frac{\sum x}{N}$

$\sum x = N \bar{X}$ = 100 x 45 = 4,500

But this is not correct $\sum x$.

Correct $\sum x$ = Incorrect $\sum x$ – wrong item + correct item

= 4,500 – 76 + 67 = 4,491

$\therefore$ Correct $\bar{X} = \frac{\text{Corrected} \sum x}{N}$

$= \frac{4491}{100} = 44.91$

MATHEMATICAL PROPERTIES

i. The sum of deviations of a set of items from this arithmetic mean (taking signs into account) is always zero, i.e., $\sum(X - \bar{X}) = 0$.

ii. The sum of the squared deviations of the items from arithmetic mean is less than the sum of the squared deviations of the items from any other value, i.e. $\sum(X - \bar{X})^2$ is less than $\sum(X - A)^2$ where, A is any other point, different from $\bar{X}$.

iii. Since, $\bar{X} = \sum X/N$, $(N\ \bar{X}) = \sum X$.

iv. If we have the arithmetic mean and number of items of two or more than two groups, we can compute **combined average** of these groups, by applying the following formula:

$\bar{X}_{12} = [(N_1 \bar{X}_1 + (N_2 \bar{X}_2)]/(N_1 + N_2)$

where,

$\bar{X}_{12}$ = Combined mean of the two groups.

$\bar{X}_1$ = Arithmetic mean of the first group.

$\bar{X}_2$ = Arithmetic mean of the second group.

N_1 = No. of items of the first group.

N_2 = No. of items of the second group.

Illustration 3

The mean weight of 20 male workers in a factory is 62 kgs and the mean weight of 30 female workers in the same factory is 53 kgs. Find out the combined mean weight of 50 workers in the factory.

Solution

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$N_1 = 20,\ \bar{X}_1 = 62,\ N_2 = 30,\ \bar{X}_2 = 53$

$$= \frac{(20\times62)+(30\times53)}{20+30} = \frac{1240+1590}{50} = 56.6 \text{ kgs.}$$

Thus, the combined mean weight of 50 workers is 56.6 kgs.

Uses: Arithmetic mean is widely used because of the following reasons:

i. Mean is the simplest average to understand and easy to compute.

ii. It is relatively reliable in the sense that it does not vary too much when repeated samples are taken from one and the same population, at least not as much as some other kind of statistical descriptions.

iii. The mean is typical in the sense that it is the center of gravity balancing the values on either side of it.

Abuses: The calculations of arithmetic mean may be simple and foolproof, but the application of the result may not be so foolproof. An arithmetic mean may not merely lack significance; it may well be positively misleading. Mean should never be accepted as significant without supporting credentials.

Example

The following table shows the result of Dyes and Pigments industry over the two years 2006-2007 and 2007-2008.

Company	Return on Net Worth (%)	
	2006-2007	2007-2008
Atul Products	24.8	14.6
Color Chem	16.7	18.9
Vanavil Dyes & Chemicals	36.5	18.9
IDI	25.9	18.7
Sudarshan Chemicals	14.0	14.4
Hanif Products	14.1	46.5
Total	132.0	132.0
Mean	22.0	22.0

If one accepts the two-yearly means, showing that no changes have occurred between the two years, there is a surprise awaiting anyone who looks back at the company details from which the means were derived.

WEIGHTED ARITHMETIC MEAN

Another aspect to be considered is the importance we assign to each observation. The arithmetic mean as we calculated it so far gives equal importance to every observation. Hence, it is called Simple Arithmetic Mean. However, it may be necessary to give different weightages or importance to different observations.

Weighted Arithmetic Mean may be defined as the average whose component items are being multiplied by certain values known as 'weights' and the aggregate of the multiplied results are being divided by the total sum of their 'weights' instead of the number of the items.

The term 'weight' stands for the relative importance of the differing items. The formula for computing weighted arithmetic mean is

$$\overline{X}_W = \frac{\sum WX}{\sum W}$$

Where,

$\overline{X}_W$ = represents the weighted arithmetic mean; and

X = represents the variable value, i.e. X_1, X_2...X_n.

Example

Comment on the performance of the companies under the ten different industries given below using weighted averages:

(Rs. in crore)

Industry	Net Profit 2007-2008 (X)	No. of Companies (W)
Granite	4.72	4
Ferro Alloys	6.48	5
Leather and Allied Products	9.22	2
Industrial Explosives	0.17	4
Hotels	22.80	9
Engineering Heavy	40.24	10
Vanaspati	5.05	5
Textiles Woollens/Worsted	38.23	6
Controlling/Construction	23.79	15
Textiles (Cotton/Blended/Yarn/Fabrics)	20.53	6
Total	171.23	66

If we ignore the number of companies under individual industry and give equal weight to all the industries, the arithmetic mean will be

$$\overline{X} = \frac{4.72 + 6.48 + 9.22 + 0.17 + 22.80 + 40.24 + 5.05 + 38.23 + 23.79 + 20.53}{10}$$

$$= \frac{171.23}{10} = \text{Rs.17.123 crore}$$

Thus, the average profit is Rs.17.123 crore. However, it may be observed that Leather and Allied Products industry contains the minimum of two companies and construction industry has as many as 15 companies. Under such circumstances, it is not desirable to give equal weights to every industry. Due weightage has to be given to the number of companies under each industry. So, the average profit can be calculated using weighted average.

$$\text{The weighted average profit per industry} = \frac{\sum WX}{\sum W}$$

$$= \frac{\left[\begin{array}{l} 4.72 \text{ x } 4 + 6.48 \text{ x } 5 + 9.22 \text{ x } 2 + 0.17 \text{ x } 4 + 22.8 \text{ x } 9 \\ + 40.24 \text{ x } 10 + 5.05 \text{ x } 5 + 38.23 \text{ x } 6 + 23.79 \text{ x } 15 + 20.53 \text{ x } 6 \end{array}\right]}{66}$$

$$= \frac{1412.66}{66} = 21.404$$

The average profit of Rs.21.404 crore is a better representative of the given data compared to the value of Rs.17.123 crore.

Median

The median, as the name suggests, is the middle value of a series arranged in any orders of magnitude i.e. ascending or descending order.

As distinct from the arithmetic mean, which is calculated from the value of every item in the series, the median is what is called a positional average. The median is just the 50th percentile value below, which 50% of the values in the sample fall. The object of median is, therefore, not only to fix a value that shall be representative of a set, but also to establish a dividing line separating the higher from the lower values.

CALCULATION OF MEDIAN

Like Arithmetic Mean, the calculation of median can be done in the case of individual series, discrete series and continuous series. This is discussed below:

Individual Series

If the data set contains an odd number of items, the middle item of the array is the median. It is mathematically represented by

Size of $\frac{n+1}{2}$ th item.

If there is an even number of items, the median is the average of the two items i.e., when the total of the frequencies is even, say, 2n, then $\frac{n}{2}$ th item, and $\frac{n}{2}+1^{th}$ item are two central items and the arithmetic mean of these two items gives the median and is mathematically represented by

$$\text{Size of } \frac{\frac{n}{2} + \frac{n}{2} + 1}{2} = \frac{\frac{n+n}{2} + 1}{2}$$

$$\text{Size of} = \frac{\frac{2n}{2}+1}{2} \text{ (cancelling 2 in numerator and denominator)}$$

we get,

$$\text{Size of } \frac{n+1}{2}$$

Note: 'n' = no. of observations in individual series, and 'N' = sum of all frequencies in discrete and continuous series, and 'N' can be substituted with 'n' in the case of discrete and continuous series.

Illustration 4

The following data relates to the sales figures of certain companies relating to the year 2000-2001:

Companies	Sales
ACC	1520
Andhra Valley	436
Excel Inds	228
Indian Hotels	239
Tata Hydro	292
Tata Power	734
Tata Tea	412
Voltas	980
Tomco	312
Tinplate Co.	256

Solution

The median for the above data can be obtained as follows:

The series should first be arranged in an appropriate order. In the present case, it is in descending order.

Company	Sales in Descending Order
ACC	1,520
Voltas	980
Tata Power	734
Andhra Valley	436
Tata Tea	412
Tomco	312
Tata Hydro	292
Tinplate Co.	256
Indian Hotels	239
Excel Inds	228

Since n = 10 i.e., even

$$\text{Median} = \text{Size of } \frac{n + 1\text{th item}}{2}$$

$$= \text{Size of } \frac{10 + 1\text{th item}}{2}$$

= Size of 5.5 items.

Therefore, median is the mean of the 5th and the 6th items, i.e. (412 + 312)/2 = 362.

Thus, the median sales value of the ten companies is 362.

Discrete Series

The same rule as of individual series regarding 'even number of variable' is applied here. The only difference between both the series is the computation of cumulative frequency.

Illustration 5

From the following data find the value of median:

Income (Rs.)	No. of Persons
2,000	48
3,000	52
1,600	32
4,000	40
5,000	12
3,600	60

Solution

Income Arranged in Ascending Order	No. of Persons (f)	c.f.
1,600	32	32
2,000	48	80
3,000	52	132
3,600	60	192
4,000	40	232
5,000	12	244

$$\text{Median} = \text{Size of } \frac{N+1}{2}\text{th item} = \frac{244+1}{2} = 122.5\text{th item}$$

Size of 122.5th item falls in the third category. Hence, median income = Rs.3,000

Continuous Series

In order to find the median, the median class is to be first located and then interpolation is to be used by assuming that items are evenly spaced over the entire class interval. The formula used for the calculation of median is

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

Where,

L = Lower limit of the median class.

f = Frequency of the median class.

cf = Cumulative frequency of the class preceding the median class.

i = Width of the class interval.

N = Total frequency.

Illustration 6

Let us find median for the following data.

Gross profit as a percentage of sales	0-10	10-20	20-30	30-40	40-50
No. of companies	21	32	43	34	23

Solution

Gross Profit (%)	No. of Companies (f)	Cumulative Frequency (cf)
0-10	21	21
10-20	32	53
20-30	43	96
30-40	34	130
40-50	23	153
	$\sum f = N = 153$	

Solution

Here, the total frequency N = 153. Median is the size of the $\frac{N}{2}$ th item, i.e. $\frac{153}{2}$ th item, i.e. the size of the 76.5th item. It lies in the class 20-30. Hence, 20-30 is the median class, of which the lower limit is 20.

Thus, L = 20, N = 153, cf = 53, f = 43, i =10

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

$$= 20 + \frac{\frac{153}{2} - 53}{43} \times 10 = 20 + \frac{76.5 - 53}{43} \times 10 = 20 + 5.46 = 25.46$$

Thus, 25.46 is the median gross profit percentage of the companies.

Calculation of Median when Class Intervals are Unequal

When the class intervals are unequal, the frequencies need not be adjusted to make the class intervals equal and the same formula for interpretation can be applied as given below:

Illustration 7

Calculate median from the following data:

Marks	0-10	10-30	30-60	60-80	80-90
No. of Students	5	15	30	8	2

Calculation of Median

Marks	f	cf
0-10	10	10
10-30	30	40
30-60	60	100
60-80	16	116
80-90	4	120

Median = Size of $\frac{N}{2}$ th item = $\frac{120}{2}$ = 60th item

Median lies in the class 30-60.

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

$L = 30,\ \frac{N}{2} = 60,\ cf = 40,\ f = 60,\ I = 30$

$$\text{Median} = 30 + \frac{60 - 40}{60} \times 30 = 40.$$

MATHEMATICAL PROPERTY

The sum of the deviations of the items from median, ignoring signs, is the least. For example, the median of 6, 10, 14, 18 and 22 is 14. The deviations from 14 ignoring signs are 8, 4, 0, 4 and 8 and the total is 24. This total is smaller than the one we obtain, if deviations are taken from any other value. Thus, if deviations are taken from 12, values ignoring signs would be 6, 2, 2, 6 and 10 and the total would be 26.

ADVANTAGES

i. It is especially useful in the case of open-end classes, since only the position and not the values of items must be known. The median is also recommended if the distribution has unequal classes, since it is easier to compute than the mean.

ii. It is not influenced by the magnitude of extreme deviations from it. For example, the median of 10, 20, 30, 40 and 150 would be 30 whereas the mean is 50. Hence, very often when extreme values are present in a set of observations, the median is a more satisfactory measure of the central tendency than the mean.

iii. In markedly skewed distributions, such as income distributions or price distributions where the arithmetic mean would be distorted by extreme values, the median is especially useful. Consequently, the median income for some purposes can be regarded as a more representative figure, for half the income earners must be receiving at least the median income.

iv. It is the most appropriate average in dealing with qualitative data, i.e. where ranks are given or there are other types of items that are not counted or measured, but are scored.

v. The value of median can be determined graphically, whereas the value of mean cannot be graphically ascertained.

vi. Perhaps the greatest advantage of median is, however, the fact that the median actually does indicate what many people incorrectly believe the arithmetic mean indicates. The median indicates the value of the middle item in the distribution. This is a clear-cut meaning and makes the median a measure that can be easily explained.

DISADVANTAGES

i. For calculating median, it is necessary to arrange the data; other averages do not need any arrangement.

ii. Since it is a positional average, its value is not determined by each and every observation.

iii. It is not capable of algebraic treatment. For example, median cannot be used for determining the combined median of two or more groups as is possible in the case of mean. Similarly, the median wage of a skewed distribution times the number of workers and will not give the total payroll. Because of this limitation, the median is much less popular as compared to the arithmetic mean.

iv. The value of median is affected more by sampling fluctuations than the value of arithmetic mean.

v. The median, in some cases, cannot be computed exactly as can the mean. When the number of items included in a series of data is even, the median is determined approximately as the mid-point of the two middle items.

vi. It is erratic, if the number of items is small.

Mode

The mode is the value which occurs most frequently in a set of observations. It is also a positional average.

CALCULATION OF MODE

Individual Series

Illustration 8

The following data relates to the share price quotations of Reliance Industries Ltd. quoted in the first fortnight of July, 2000:

189, 197.50, 193.75, 188.75, 193.75, 217.50, 207.50, 193.75, 210.50, 193.75, 182.50, 183.75, 191.25, 193.75, 188.75

Solution

Let us calculate the mode for the above data.

Price Quotation	No. of Times it Occurs
189.00	1
197.50	1
193.75	5
188.75	2
217.50	1
207.50	1
210.50	1
182.50	1
183.75	1
191.25	1
Total	15

Since the share price 193.75 has occurred maximum number of times, the mode is 193.75.

Thus, the process of determining mode in the case of individual observations essentially involves grouping of data.

Discrete Series

Where the mode is determined just by inspection, an error of maximum frequency and the frequency preceding it, on succeeding it, is very small and the items are heavily concentrated on either side. In such cases, it is desirable to prepare a grouping table and an analysis table. These tables help us in ascertaining the modal class.

A grouping table has six columns. In column 1, the maximum frequency is marked on put in a circle; in column 2 frequencies are grouped in two's; in column 3 leave the first frequency and then group the remaining in two's; in column 4 group the frequencies in three's; in column 5 leave the first frequency and group the frequencies in three's; and in column 6 leave the first two frequencies and then group the remaining in three's. In each of these cases take the maximum total and mark it in a circle on by bold type.

After preparing the grouping table, prepare an analysis table. While preparing this table, put column number on the left-hand side and the various probable valves of mode on the right-hand side. The values against which frequencies are the highest are marked in the grouping table and then entered by means of a bar in the relevant 'box' corresponding to the values they represent.

Illustration 9

Calculate the value of mode for the following data:

Marks	10	20	30	40	50	60	70
Frequency	8	12	35	32	27	15	10

Solution

Calculation of Mode

X	F	II	III	IV	V	VI
10	8					
		20				
20	12		47	55	79	
30	35					
		67				94
40	32		59	74	52	
50	27					
		42				
60	15		27			
70	12					

Analysis Table

Col.No.	30	40	50
I	1	–	–
II	1	1	–
III	–	1	1
IV	–	1	1
V	1	1	–
VI	1	1	1
	4	5	3

Corresponding to the maximum total 5, the value of the variable is 40. Hence modal value is 40.

Continuous Series

For calculating mode from a frequency distribution, the following formulas can be used.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

Where,

L = lower limit of the modal class;

Δ_1 = difference between the frequency of the modal class and the frequency of the pre-modal class, i.e., preceding class; and

Δ_2 = difference between the frequency of the modal class and the frequency of the post-modal class, i.e., succeeding class;

i = the class interval of the modal class.

Another form of this formula is

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Where,

L = Lower limit of the modal class which is the class having the maximum frequency.

f_0, f_2 = Frequencies of the classes preceding and succeeding the modal class respectively.

f_1 = Frequency of the modal class.

i = Class interval.

Illustration 10

Let us find the mode for the data given below:

Sales in Crore	0-8	8-16	16-24	24-32	32-40
No. of companies	19	25	36	43	28

Solution

Here, the largest frequency is 43, and it lies in the class 24-32. So, the modal class is 24-32. Therefore, L = 24, f_2 = 28, f_1 = 36, i = 8, f = 43

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 24 + \frac{43-36}{2 \times 43 - 36 - 28} \times 8$$

$$= 24 + 2.55 = 26.55$$

It is to be noted that if the final class has been recognized as a class with the highest frequency, it is regarded that 'mode' is ill-defined.

Example

Calculate Karl Pearson's coefficient of skewness from the data given below:

Weekly Wages (Rs.)	No. of Workers (f)
30-40	5
40-50	6
50-60	8
60-70	10
70-80	25
80-90	30
90-100	36
100-110	50
110-120	60
120-130	70

Solution

It is found that the 120-130 has highest frequency. In such case, mode cannot be defined.

GRAPHICAL CALCULATION

For calculating the mode of the grouped data graphically, the following procedure is adopted:

i. Draw a histogram of the data; the modal class is the tallest rectangle.

ii. Draw a line from the top right corner of the tallest rectangle to the top right corner of the preceding rectangle.

iii. Draw a line from the top left corner of the tallest rectangle to the top left corner of the succeeding rectangle.

iv. Draw a line perpendicular to the X-axis from the point of intersection of lines drawn in steps 2 and 3. The point of intersection of the perpendicular line with the X-axis represents the mode.

Illustration 11

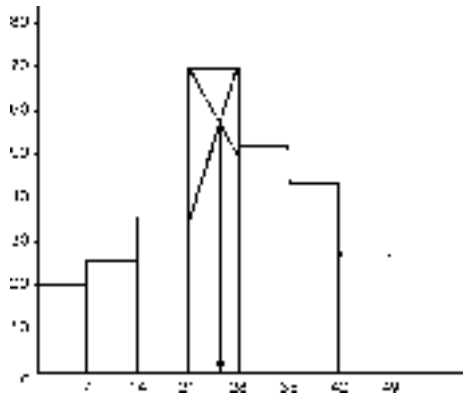
Let us calculate the Mode using the graphical method for the following distributions:

Gross Profits as Percentage of Sales	0-7	7-14	14-21	21-28	28-35	36-42	42-49
Number of Companies	19	25	36	72	51	43	28

Solution

Using the procedure described above, we draw the histogram and other lines for the calculation of mode. The mode is 25.

Figure



BIMODAL DISTRIBUTION – CALCULATION OF MODE

There may be two values that occur with the same maximum frequency. Such distribution is called bimodal. In a bimodal distribution, the value of mode cannot be determined with the help of the formula given above. In such instances where a distribution is bimodal and nothing can be done to change it, the mode should not be used as a measure of central tendency. If we have more points of maxima for the frequency distribution, the distribution is known as multi-modal distribution.

EMPIRICAL MODE

Where mode is ill-defined, its value may be ascertained by the following formula based upon the empirical relationship between Mean, Median and Mode:

$$\text{Mode} = 3\ \text{Median} - 2\ \text{Mean}$$

This measure is called the empirical mode.

Illustration 12

The following is the data pertaining to the marks of 50 students.

Marks	No. of Students
30	7
35	13
40	15
45	10
50	5

You are required to calculate mode based on the empirical method.

Solution

To calculate mode on the basis of the empirical method we need to calculate arithmetic mean and median:

Arithmetic Mean

$$= \frac{\Sigma fx}{N} = \frac{1{,}965}{50} = 39.3$$

Median

$$= \text{Size of } \frac{N + 1^{th} \text{ item}}{2}$$

$$= \frac{50+1}{2} = 25.5\text{th item } \frac{50+1}{2} = 25.5\text{th item and}$$

falls in the third category of marks = 40

Mode

= 3 Median – 2 Mean

$= (40 \times 3) - (2 \times 39.3) = 120 - 78.6 = 41.4$

Advantages

i. By definition, mode is the most typical or representative value of a distribution. Hence, when we talk of modal wage, modal size of shoe or modal size of family it is this average that we refer to. The mode is a measure, which actually does indicate what many people incorrectly believe the arithmetic mean indicates. The mode is the most frequently occurring value. If the modal wage in a factory is Rs.2,000, more workers receive Rs.2,000 than any other wage. This is what many believe the "average" wage always indicates, but actually such a meaning is indicated only if the average used is the mode.

ii. It is not affected by extremely large or small items. For example, the mode of values 1, 4, 4 and 10 is 4 and the mode of values 1, 4, 4 and 1,000 is also 4.

iii. Its value can be determined in open-end distributions without ascertaining the class limits.

iv. It can be used to describe qualitative phenomenon. For example, if we want to compare the consumer preferences for different types of products, say, soap, toothpastes, etc. or different media of advertising, we should complete the modal preferences expressed by different groups of people.

v. The value of mode can also be determined graphically, whereas the value of mean cannot be graphically ascertained.

Disadvantages

i. The value of mode cannot always be determined. In some cases we may have a bimodal series.

ii. It is not capable of algebraic manipulations. For example, from the modes of two sets of data we cannot calculate the overall mode of the combined data. Similarly, the modal wage times the number of workers will not give the total payroll except, of course, when the distribution is normal and then the mean, median and mode are all equal.

iii. The value of mode is not based on each and every item of the series.

iv. It is not a rigidly defined measure. There are several formulae for calculating the mode, all of which usually give somewhat different answers. In fact, mode is the most unstable average and its value is difficult to determine.

v. While dealing with quantitative data, the disadvantages of the mode outweigh its good features and hence it is seldom used.

APPROPRIATENESS OF THE THREE PRINCIPAL AVERAGES

There is no single answer to the question: Which average is to be used to describe statistical data?

There are situations where none of the three averages is fully satisfactory. For example, if the number of items in a series is very small, none of these averages has any meaning as a measure of central tendency. Moreover, if the items are what has been called "freakishly deployed" (as when concentrated at one end), an average for the series can be found, but it is not descriptive of the series.

The data available may exclude the use of certain averages. For instance, in open-end distributions, the mean cannot be accurately found. However, the median or mode can be used unless the median or modal class happens to be open-ended. In a situation of bimodality, the mode makes no sense as a measure of central tendency.

The characteristics of the averages and the demands of the problem to be solved usually determine which one is to be used. The data may call for a certain average.

In such problems as the average score on personnel tests or the average productivity rate of workers, where the values in the series really ranks, the median is the average to use. These scores and rates are not additive, that is, they indicate not unit quantities, but rather the position of an individual with respect to other individuals. Therefore, a positional average namely the median is appropriate.

The arithmetic mean is the most commonly used and the best known of the averages, and is preferred unless precluding circumstances are present, such as extreme values at either end of the series or open-end classes, or varying class intervals, or unless we definitely wish to establish the most frequent value or some other positional average. Where further computational techniques are involved in the investigation, the mean is the average to be used.

Great care should be taken in choosing an average. On occasions, it may even be advisable to work-out all three averages and present them. The added burden is preferable to the use of a single average that may be an incomplete description.

However, considering all these points, the statistician, guided by the desire to present an accurate picture of the data and to command respect, is the final judge of, which average is the most appropriate.

Comparison of the Principal Averages

i. The mean, median, and mode are located at the same point in a symmetrical frequency distribution.

ii. The mean is a computed average, whereas the median and the mode are positional.

iii. Extreme values in the series affect the utility of the mean, but not of the median or the mode.

iv. The presence of open-end classes excludes the use of the mean, but not of the median and sometimes not of the mode.

v. Varying class intervals usually make the mean unreliable, but do not affect the median. In such a case, the mode may be found but only by reclassification through frequency densities.

vi. Mean may be combined, but not median and mode.

vii. The mean has four mathematical properties, which make it indispensable in advanced statistical work. The median has one mathematical property, while the mode has none.

viii. To compute the median and the mode, the data has to be sorted.

ix. The median and the mode may be found graphically, but not the mean.

Geometric Mean

Geometric Mean is defined as the nth root of the product of numbers to be averaged. The geometric mean of numbers X_1, X_2, X_3.....X_n is given as:

$$G = (X_1 \times X_2 \times X_3 \ldots\ldots X_n)^{1/n}.$$

It is a type of mathematical average.

ASCERTAINING THE RATE OF GROWTH OVER TIME THROUGH GEOMETRIC MEAN

The peculiar nature of growth over time is on account of compounding. This is explained with the following example:

Suppose the price of a share was Rs.100 in January, 1999. By January, 2000 it increased by 100% to Rs.200. Further, in the next one year it decreased by 50% so that in January, 2001 the price was again Rs.100. Hence, the growth rate over the two years was zero.

If we take the arithmetic mean of the annual growth rates we get

(100%) + (–50%)/2 = 25%

which is clearly incorrect because we have seen that the average rate is zero.

We can obtain the correct answer by using the geometric mean.

To calculate the geometric mean, we first convert the percentage growth rates into quantity ratios (also called growth factors).

$$\text{Quantity ratio} = \frac{\%\ \text{growth rate}}{100} + 1$$

So,

quantity ratio for 1999 = (100/100) + 1 = 2

This implies that every Re.1 in January, 1999 grew to Rs.2 at the end of the year. Hence, the price Rs.100 in January, 1999 grew to Rs.200 in one year.

Quantity ratio for 1999 = (– 50/100) + 1 = 0.5

Hence, every Re.1 in January, 2000 reduced to Re.0.50 in the next year.

Geometric mean = [Product of quantity ratios]1/n

Where,

n = Number of quantity ratios.

So,

Geometric Mean = (2 x 0.5)1/2 = 11/2 = 1

Average Growth rate = Geometric Mean –1 = 1 – 1 = 0.

Calculation of Compound Interest

The expression used in calculating compound interest formula is

$$P_n = P_0(1 + r)^n$$

Where,

P_n = The value at the end of period n.

P_0 = The value at the beginning of the period.

r = Rate of compound interest per annum (expressed as a fraction).

n = Number of years.

Illustration 13

A sum of Rs.1,000 deposited today in a bank gets doubled in a period of 6 years. Find the annual rate of interest.

Solution

Let r be the rate of interest.

Applying the formula,

$P_n = P_0(1 + r)^n$

i.e. $2000 = 1000\ (1 + r)^6$

So $(1 + r)^6 = \frac{2000}{1000} = 2$ (product of quantity ratios)

$1 + r = (2)^{1/6} = 1.1225$

$r = 0.1225 = 12.25\%$

Thus, the rate of interest per annum = 12.25%.

Uses: The geometric mean is used to find the average percent increase in sales, production, population or other economic or business series overtime.

Illustration 14

The following data relates to Voltas Ltd.

Year	Sales (Rs. in million)
2005-2006	6670.0
2006-2007	7794.6
2007-2008	9176.2

The growth rate for the year 2006-2007 = $\frac{1124.6}{6670}$ x 100 = 16.86%

The growth rate for the year 2007-2008 = $\frac{1381.6}{7794.6}$ x 100 = 17.73%

We can find that the sales of Voltas Ltd. has been increasing year by year, but at different growth rates. Now, the compounded annual growth rate can be arrived at by taking the geometric mean for the two quantity ratios.

$GM = \sqrt{1.1686 \times 1.1773} = 1.1729$

Growth rate = 1.1729 – 1 = 0.1729 or 17.29%

Thus, the compounded annual sales growth rate of Voltas Ltd., for the years 2005-06, 2006-2007 and 2007-2008 is 17.29%.

PROPERTIES OF GEOMETRIC MEAN

The following are the two important properties of geometric mean, which are based on mathematics.

The product of the value of series remains unchanged when the value of geometric mean is substituted for each individual value. For example:

2 x 4 x 8= 64 = 4 x 4 x 4

The sum of the deviations of the logarithms of the original observations above or below the logarithm of the geometric mean is equal. It means the geometric mean brings a balance between ratio deviations of the given observations of the data. For example:

(4/2) (4/4) = 2 = (8/4)

CALCULATION OF GEOMETRIC MEAN

Individual Observations

$$GM = \text{Antilog}\left(\frac{\sum \log X}{N}\right)$$

Illustration 15

Daily production of a company is given below. Find out the GM

120 143 205 318 236 379 473 395 458

Solution

Calculation of Geometric mean

Daily Production X	Log X
175	2.2430
120	2.0792
143	2.1553
205	2.3118
318	2.5024
236	2.3729
379	2.5786
473	2.6749
395	2.5966
458	2.6609
	$\sum \log x = 24.1756$

$$GM = AL\left(\frac{\sum \log X}{N}\right) = AL\left(\frac{24.1756}{10}\right) = AL\ (2.41756) = 261.6$$

Discrete Series

$$GM = \text{Antilog}\left(\frac{\sum f \log X}{N}\right)$$

Illustration 16

Find the GM for the following data

x	2	3	5	6	4
Number of Companies	10	15	18	12	7

Solution

x	f	Log x	f × Log x
2	10	0.3019	3.0100
3	15	0.4771	7.1565
5	18	0.6990	12.5820
6	12	0.7782	9.3384
4	7	0.6021	4.2147
	N = 62		$\sum f \text{ Log } x =$ 36.3016

$$GM = \text{Antilog} \left(\frac{\sum f \log X}{N} \right)$$

$$= \text{Antilog} \left(\frac{36.3016}{62} \right)$$

$$= \text{Antilog}\ (0.5855) = 3.850$$

CONTINUOUS SERIES

The geometric mean in continuous series can be calculated with the following formula:

$$GM = AL \left[\frac{\sum f \log X}{N} \right] = AL \left[N \frac{\sum f \log m}{N} \right];\ (m=X)$$

Illustration 17

Find the Geometric mean for the data given below:

Marks	Frequency
2-6	6
6-10	10
10-14	18
14-18	30
18-22	15
22-26	12
26-30	10
30-34	6
34-38	2

Solution

Marks	Mid Point (m)	Frequency (f)	Log m	f log m
2-6	4	6	0.6021	3.6126
6-10	8	10	0.9031	9.031
10-14	12	18	1.0792	19.4256
14-18	16	30	1.2041	36.123
18-22	20	15	1.3010	19.515
22-26	24	12	1.3802	16.5624
26-30	28	10	1.4472	14.472
30-34	32	6	1.5051	9.0306
34-38	36	2	1.5563	3.1126
		N = 109		$\sum f \log m$ 130.8848

$$GM = AL \left[\frac{\sum f \log X}{N} \right] = AL \left[N \frac{\sum f \log m}{N} \right];\ (m = X)$$

$$AL \left[\frac{130.8848}{109} \right] = AL\ (1.2007) = 15.88$$

WEIGHTED GEOMETRIC MEAN

Weighted geometric mean can be calculated with the help of the following formula:

$$GM_w = \text{Antilog} \left[\frac{\sum W \log X}{\sum W} \right]$$

Illustration 18

Find the weighted geometric mean from the following data:

Group	Index Number	Weights
Food	260	23
Fuel and Lighting	180	5
Clothing	220	4
House Rent	230	10
Education	120	6
Misc.	200	2

Solution

Calculation of Weighted Geometric Mean

Group	Index Number (X)	Weights (W)	Log X	W Log X
Food	260	23	2. 4150	55.545
Fuel and Lighting	180	5	2.2553	11.2765
Clothing	220	4	2.3424	9.3696
House Rent	230	10	2.3617	23.617
Education	120	6	2.0792	12.4752
Misc.	200	2	2.3010	4.602
		$\sum W = 50$		$\sum W \text{ Log } X = 116.8853$

$$GM_w = AL\left[\frac{\sum W \log}{\sum W}\right] = AL\left[\frac{116.8853}{50}\right] = AL\ (2.3377) = 217.6$$

Merits

1. It is rigidly defined.
2. It is useful in average ratios and percentages and in determining rates of increase and decrease,
3. It gives less weight to large items and more to small ones than does the arithmetic average. Because, the geometric mean is never larger than the arithmetic mean.
4. It is capable of algebraic manipulation.

Limitations

1. It is difficult to compute and to interpret and so has restricted application.
2. It cannot be computed when there are both **negative** and positive values in a series or one or more of the values are **zero**.

Harmonic Mean

The Harmonic Mean is based on the reciprocals of numbers averaged. It is defined as the reciprocal of the arithmetic mean of the reciprocal of the given individual observations. Thus, by definition

$$HM = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + ...+ \frac{1}{X_n}} = \frac{N}{\sum\left(\frac{1}{X}\right)}$$

Where,

X_1, X_2, X_3, etc., refer to values of various items of the variable. It is a mathematical average.

CALCULATION OF HARMONIC MEAN

Individual Series

$$HM = \frac{N}{\sum\left(\frac{1}{X}\right)}$$

Illustration 19

Calculate the harmonic mean from the following data:

1578, 957 438 76 0.5 0.02 0.003 0.0008

Solution

Calculation of Harmonic Mean

X	(1/X)
1578	0.0006
957	0.0010
438	0.0022
76	0.0131
0.5	2
0.02	50
0.003	333.3333
0.0008	1250
	$\sum \frac{1}{X} = 1635.3502$

$$HM = \frac{N}{\sum\left(\frac{1}{X}\right)} = \frac{8}{1635.3502} = 0.005$$

Discrete Series

$$HM = \frac{N}{\sum\left(f \times \frac{1}{X}\right)} = \frac{N}{\sum\left(\frac{f}{X}\right)}$$

Illustration 20

From the following data, compute the value of harmonic mean

Marks	10	20	30	40	50
No. of Students	15	25	30	20	10

Solution

Calculation of Harmonic Mean

Marks (X)	f	F/X
10	15	1.5
20	25	1.25
30	30	1.00
40	20	0.5
50	10	0.2
	N = 100	$\sum\left(\frac{f}{X}\right) = 4.45$

$$HM = \frac{N}{\sum\left(\frac{f}{X}\right)} = \frac{100}{4.45} = 22.47$$

Continuous Series

$$HM = \frac{N}{\sum\left(f \times \frac{1}{X}\right)} = \frac{N}{\sum\left(\frac{f}{m}\right)}$$

Illustration 21

From the following data compute the value of harmonic mean:

Class Interval	100-200	200-300	300-400	400-500	500-600
Frequency	6	8	12	9	4

Solution

Class Interval	Mid-points	Frequency	f/m
100-200	150	6	0.04
200-300	250	8	0.032
300-400	350	12	0.034
400-500	450	9	0.02
500-600	550	4	0.007
		N = 39	0.133

$$HM = \frac{N}{\sum\frac{f}{m}} = \frac{39}{0.133} = 293.234.$$

WEIGHTED HARMONIC MEAN

Weighted Harmonic Mean is calculated with the help of the following formula:

$$WHM = \frac{\sum W}{\sum (W/X)}$$

Case 1:

Consider a company consisting of only two divisions A and B. The calculation of the net profit margin for the two divisions as well as for the company as a whole is shown below:

	Division A	Division B	Whole company (A + B)
Net Profits (Rs. in crore)	6	1	7
Sales (Rs. in crore)	40	40	80
Net Profit Margin	15%	2.5%	8.75%

Here, we see that the net profit margin for the company as a whole is 8.75% which is nothing but the **Simple Arithmetic Mean** of the net profit margins of the two divisions A and B.

$$= \frac{(15+2.5)}{2} = \frac{17.5}{2} = 8.75$$

So, the simple arithmetic mean has a significance here. Note that the net profit margins of the two divisions A and B have been calculated with the same denominator (Sales = Rs.40 crore). In general, we can say that the appropriate mean for a set of ratios which have been calculated with the same denominators is the simple arithmetic mean.

Case 2:

Consider a company consisting of only two divisions, X and Y. The calculation of the net profit margins for the two divisions as well as for the company as a whole is given below:

	Division X	Division Y	Whole company (X + Y)
Net Profits (Rs. in crore)	6	6	12
Sales (Rs. in crore)	40	80	120
Net Profit Margin	15%	7.5%	10%

Here, we find that the simple arithmetic mean of the net profit margins of the two divisions is (15 + 7.5)/2 = 22.5/2 = 11.25%. This is not equal to the net profit margin for the whole company which is 10%.

Let us now find the simple harmonic mean of the divisional net profit margins. The Simple Harmonic Mean (SHM) is given by $N/\left(\sum\frac{1}{x}\right)$ where the X's are the divisional net profit margins. So, SHM = $2/\left(\frac{1}{15}+\frac{1}{7.5}\right)$ = 10%

So, here, the net profit margin for the whole company is the simple harmonic mean of the net profit margins of the company's two divisions.

In general, we can say that the appropriate mean for a set of ratios which have been calculated with the same numerators is the simple harmonic mean.

Case 3:

Consider a company consisting of only two divisions P and Q. Below we have the calculation of the net profit margins for the two divisions as well as for the company as a whole.

	Division P	Division Q	Whole company
Net Profits (Rs. in crore)	6	1	7
Sales (Rs. in crore)	50	20	70
Net Profit Margin	12%	5%	10%

Here, we find that the simple arithmetic mean of the net profit margins of the two divisions is (12 + 5)/2 = 8.5%. The simple harmonic mean of the net profit margins of the two divisions is $= \frac{2}{\frac{1}{12}+\frac{1}{5}} = 7.06\%$.

The net profit margin for the company is equal to neither of the above means. So, which is the appropriate mean in this case?

In this case, where the ratios have been calculated with neither the same numerators nor the same denominators, the appropriate mean is the 'Weighted Mean'. But which weighted mean?

The choice of weighted mean would depend on the weights we use. If we use the numerators (Net Profits) as weights, the appropriate mean is the Weighted Harmonic Mean (WHM). So,

$$\text{WHM} = \frac{6+1}{\frac{6}{12}+\frac{1}{5}} = 10\%$$

which is the net profit margin for the company as a whole.

On the other hand, if we use the denominators (Sales) as weights, the appropriate mean is the Weighted Arithmetic Mean (WAM).

So,

$$\text{WAM} = \frac{50 \text{ x } 12 + 20 \text{ x } 5}{50 + 20} = 10\%$$

Which is the net profit margin for the company as a whole.

Hence, we see that the harmonic mean plays a complementary role to the arithmetic mean enabling us to arrive at the same results by alternate methods. Although we have considered a specific ratio (Net Profit Margin), the logic applies to any other ratio as well. We have considered the mean of only two ratios, but the logic could be extended to any number of ratios.

We list the rules given above:

1. The appropriate mean for a set of ratios which have been calculated with the same denominators is the simple arithmetic mean.
2. The appropriate mean for a set of ratios which have been calculated with the same numerators is the simple harmonic mean.
3. The appropriate mean for a set of ratios using the denominators of the ratio data as weights is the weighted arithmetic mean.
4. The appropriate mean for a set of ratios using the numerators of the ratio data as weights is the weighted harmonic mean.

Remark:

The above analysis is valid only for a specific type of ratio x/y where the average ratio is given by $\sum x / \sum y$. This kind of ratio is encountered when the ratio for a company as a whole $\sum x / \sum y$ is split up into divisional or departmental ratios x/y.

The above analysis is not valid for ratios covering growth rates over long periods of time where the appropriate mean is the Geometric Mean.

Uses:

It is used in special cases where the computation is required to find the average rate of increase in losses of a public limited concern or the average price at which the goods have been sold or the speed at which the journey has been transformed.

Merits

a. It is rigid and definite.

b. It is based on all the observations of the given data.

c. It is capable for further algebric treatment

d. It is not affected by sampling fluctuations.

e. It is apt for calculating time and rate related aspects of the data.

Limitations

a. It is difficult to understand the computation of harmonic mean.

b. Its calculation involves complexity as it is based on reciprocals.

c. It gives undue weights to small items and ignores bigger items.

d. In the case of zero or negative values, it cannot be computed.

RELATIONSHIP AMONG THE MATHEMATICAL AVERAGES

In any distribution when the original items differ in size, the value of Arithmetic Mean (AM), Geometric Mean (GM), and Harmonic Mean (HM) would also differ and will be in the following manner:

AM ≥ GM ≥ HM i.e., AM is greater than GM and Gm is greater than HM and all of them will be equal if the values of a variable are identical. This can be explained below:

Let a and b be two positive quantities such that $a \neq b$.

Then AM and HM of these quantities are

$$AM\left(\overline{X}\right) = \frac{a+b}{2};\ GM\sqrt{ab};\ HM = \frac{2ab}{a+b}$$

As discussed above $\frac{a+b}{2} > \sqrt{ab}$ or $a+b > 2\sqrt{ab}$

$$= a+b-2\sqrt{ab} = 0$$

$$= \left(\sqrt{a}-\sqrt{b}\right)^2 > 0 \qquad \left(Q\ a+b-2\sqrt{ab} = \left(\sqrt{a}-\sqrt{b}\right)^2\right)$$

But the square of any real quantity is positive. Hence, $\left(\sqrt{a}-\sqrt{b}\right)^2$ will be positive. Hence, $\frac{a+b}{2} > \sqrt{ab}$

Let us now prove that GM > HM

i.e., $\sqrt{ab} > \frac{2ab}{a+b}$ or $a+b > 2\sqrt{ab}$ as stated above.

Hence, GM > HM.

From the above discussion it is clear that, AM > GM and GM > HM. Hence, It is obvious that AM > GM > HM.

CHOICE OF A SUITABLE AVERAGE

All types of averages are apt and can be applied for different purpose. However, their application differs according to different circumstances. The following points are worth noting while selecting an appropriate average:

- The purpose for which the choice is to be made must be taken care of. For example, if the purpose is to give all the items in the series due importance, arithmetic mean is a suitable average.
- The form and the nature of the data also plays a vital role in this respect. For example, if the data is a skewed data, mode or median will be apt for the data.
- The choice of a suitable average also depends on its adaptability to further algebraic treatment. If the average is to be used for further algebraic treatment, arithmetic average holds good.
- If the data is qualitative in nature, median holds good.

Thus, the above factors determine the selection of a suitable average.

ADDITIONAL ILLUSTRATIONS

Illustration 1

The average monthly salary of the supervisors and workmen of Alpha Ltd. is Rs.6,000. The average monthly salary of the supervisors is Rs.9,000 and the average monthly salary of the workmen is Rs.5000. Moreover, the average monthly salary of the middle level managers of the company is Rs.14,000. What is the percentage of supervisors among the total number of workers and supervisors in the company?

Solution

Let N_1 denote the number of supervisors and N denote the total number of supervisors and workmen

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + (N - N_1)\bar{X}_2}{N} \text{ or}$$

$$\frac{9000N_1 + 5000(N - N_1)}{N} = 6000 \text{ or}$$

$$\frac{9000N_1 + 5000N - 5000N_1}{N} = 6000 \text{ or}$$

$$\frac{4000N_1 + 5000N}{N} = 6000 \text{ or}$$

$$\frac{4000N_1}{N} + 5000 = 6000 \text{ or}$$

$$\frac{4000N_1}{N} = 1000 \text{ or}$$

$$\frac{N_1}{N} = \frac{1000}{4000} = \frac{1}{4} = 0.25$$

$\therefore$ Percentage of supervisors = $0.25 \times 100 = 25\%$.

Illustration 2

From the following data compute the mean and median:

No. of Shares	2000-2500	2500-3000	3000-3500	3500-4000	4000-4500	4500-5000
Frequency ('000)	21	18.5	16	14	10.50	3

Solution

Calculation of Mean

No. of Shares	Mid Value(x)	Frequency ('000) (f)	(2) × (3) ('000)
(1)	(2)	(3)	(4)
2000-2500	2250	21.00	47,250
2500-3000	2750	18.50	50,875
3000-3500	3250	16.00	52,000
3500-4000	3750	14.00	52,500
4000-4500	4250	10.50	44,625
4500-5000	4750	3.00	14,250
		$\sum f = 83$	$\sum fx = 2,61,500$

Mean number of applications

$$= \frac{\sum fx}{N} = \frac{2,61,500 \times 1,000}{83 \times 1,000} = 3,150.6$$

Calculation of Median

No. of Shares	Frequency	Cumulative Frequency
	('000)	('000)
2000 – 2500	21.00	21.00
2500 – 3000	18.50	39.50
3000 – 3500	16.00	55.50
3500 – 4000	14.00	69.50
4000 – 4500	10.50	80.00
4500 – 5000	3.00	83.00

Median position = $\frac{N}{2} = \frac{83,000}{2} = 41,500$

$\therefore$ Median class is 3000 – 3500.

$\therefore$ Median = $L + \frac{N/2 - cf}{f} \times i$

$\therefore$ Median number of applications = $3000 + \frac{41,500 - 39,501}{16,000} \times 500 = 3,062.5$

Illustration 3

A data set includes some quantities. The sum of reciprocals of the quantities in the data set is $\frac{7}{8}$. The harmonic mean of the data set is $3\frac{3}{7}$. The data set is expanded by including the quantities 5 and 10 into it. What is the harmonic mean of the expanded data set?

Solution

Harmonic mean of a set of 'n' quantities, $H = \frac{n}{\sum \frac{1}{x_i}}$

Given: $\sum \frac{1}{x_i} = \frac{7}{8}$ $\quad H = 3\frac{3}{7} = \frac{24}{7}$

From above, $n = H. \sum \frac{1}{x_i}$

$\therefore \quad n = \frac{24}{7} \times \frac{7}{8} = 3$

Harmonic mean after the expansion $= \frac{n+2}{\frac{7}{8}+\frac{1}{5}+\frac{1}{10}} = \frac{3+2}{\frac{47}{40}} = 4\frac{12}{47}$

Illustration 4

Calculate the arithmetic mean from the following data

Marks	No. of Students
Less than 80	100
70	90
60	80
50	60
40	32
30	20
20	13
10	5

Solution

Calculation of Arithmetic Mean

Marks	Frequency	M	(X – 35)/10 = d	fd
0-10	5	5	– 3	– 15
10-20	8	15	– 2	– 16
20-30	7	25	– 1	– 7
30-40	12	35	0	0
40-50	28	45	+1	+28
50-60	20	55	+2	+40
60-70	10	65	+3	30
70-80	10	75	+4	+40
	N = 100			$\sum$ fd = 100

$\bar{X} = A + \frac{\sum fd}{N} \times c = 35 + \frac{100}{100} \times 10 = 45$

Illustration 5

Calculate the median and quartiles for the following data:

Marks	No. of Students
5-10	6
10-15	5
15-20	15
20-25	20
25-30	10
30-35	4
35-40	2

Solution

Calculation of Median and Quartile

Marks	No. of Students	cf
5-10	6	6
10-15	5	11
15-20	15	26
20-25	20	46
25-30	10	56
30-35	4	60
35-40	2	62

$$\text{Median} = \frac{N}{2} = \frac{62}{2} = 31$$

$\therefore$ Median class is 20-25

$$\therefore \text{Median} = L + \frac{N/2 - cf}{f} \times i$$

$$\therefore \text{Median number of applications} = 20 + \frac{31-26}{20} \times 5 = 21.25$$

$$Q_1 = \frac{N}{4} = \frac{62}{4} = 15.5$$

$\therefore$ Quartile class is 15-20

$$\therefore \quad Q_1 = L + \frac{N/4 - cf}{f} \times i$$

$$\therefore \quad Q_1 = 15 + \frac{15.5-11}{15} \times 5 = 16.5$$

$$Q_3 = \frac{3N}{4} = \frac{3 \times 62}{4} = 46.5$$

$\therefore$ Quartile class is 25-30

$$\therefore \quad Q_3 = L + \frac{3N/4 - cf}{f} \times i$$

$$\therefore \quad Q_3 = 25 + \frac{46.5-46}{10} \times 5 = 25.25$$

Illustration 6

Find the missing frequency for the following data when mean = 34

Marks	No. of Students
0-10	5
10-20	15
20-30	20
30-40	?
40-50	20
50-60	10

Solution

Marks	No. of Students	m	fm
0-10	5	5	25
10-20	15	15	225
20-30	20	25	500
30-40	?	35	35x
40-50	20	45	900
50-60	10	55	550
	N = 70 + X		2200 + 35x

$$\bar{X} = \frac{\sum fm}{N}$$

$$34 = \frac{2200 + 35x}{70 + x}$$

$34 (70 + x) = 2200 + 35x$

$2380 + 34x = 2200 + 35x$

$34x - 35x = 2200 - 2380$

$-x = -180$

on $x = 180$.

SUMMARY

- Statistical data helps in arriving at a single value which represents the entire data and since this single value (called the average) has a tendency to be somewhere at the center and within the range of all values, it is also known as the measure of central tendency. The average value is a representative of all the items and the prerequisites of a good average are that it should be well defined, easy to compute, dependent on all the observed values, should fluctuate relatively less and should be capable of mathematical manipulation.
- There are three mathematical averages (Arithmetic Mean, Geometric Mean, and Harmonic Mean) and two positional averages (Median, and Mode).
- Arithmetic Mean can be computed by dividing the sum of all observations by the number of observations. Weighted Arithmetic Mean may be defined as the average whose component items are being multiplied by certain values known as 'Weights' instead of the sum of the items.

- Median is the middle value of a series arranged in any of the orders of magnitude (ascending or descending order). Mode is the value which occurs most frequently in a set of observations on the point of maximum frequency and around which other items of the set cluster densely.
- Geometric mean is defined as the nth root of the product of numbers to be averaged and is used to find the percentage increase in sales, production or any other economic or business series over time.
- Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the given individual observations.
- All averages are used extensively and have certain merits. However, they are not free from limitations.
- Apart from the measures mentioned above, there are certain positional measures which divide a series into equal parts. Example quartiles, deciles and percentiles.
- There exists a relationship between the three mathematical averages i.e., AM, GM, and HM.
- Certain factors play an important role in selecting an appropriate measure.

Chapter VII

Measures of Dispersion

After reading this chapter, you will be conversant with:

- Meaning of Dispersion
- Properties of a Good Measure of Variation/Dispersion
- Significance of Measuring Variation/Dispersion
- Methods of Studying Variation/Dispersion
- Relationship between Quartile Deviation, Standard Deviation, and Mean Deviation
- Graphical Method – Lorenz Curve
- Additional Illustrations

Introduction

Averages or the Measures of Central Tendency discussed in the previous chapter deals with the concentration of the observation around the central part of the distribution. However, the averages have their own limitations. A single value cannot describe a set of observations unless all the observations are similar. Such an average cannot give the complete idea about the distribution because the various observations may have same average but differ widely from each other in various ways.

In the words of George Simpson and Fritz Kafka "An average does not tell the full story. It is hardly fully representative of a mass unless we know the manner in which the individual items scatter around it. A further description of the series is necessary if we are to gauge how representative the average is."

So, we need other measures which support and supplement the Measures of Central Tendency known as Measures of Dispersion.

MEANING OF DISPERSION

In literal sense, Dispersion mean 'Scatteredness'. It measures the extent to which the items vary from the central value. It is also known as Average of the Second order since it gives average of the differences of various items from the average. Dispersion gives an idea of the homogeneity or heterogeneity of the distribution.

Definitions

According to A.L. Bowley "Dispersion is the measure of the variations of the items."

According to Brook and Dick "Dispersion or spread is the degree of the scatter or variation of the variables about a central value."

Spiegel defines dispersion as "the degree to which numerical data tend to spread about an average value is called the variation of dispersion of the data."

"The measurement of the scatterness of the mass of figures in a series about an average is called measure of variation or dispersion."

PROPERTIES OF A GOOD MEASURE OF VARIATION/DISPERSION

A measure of dispersion with the following properties is considered a good measure:

i. It should be simple to understand and easy to compute.

ii. It should be defined rigidly.

iii. It should be based on each and every items of distribution.

iv. Extreme items should not be affected.

v. It should have sampling stability.

vi. Further algebraic treatment is amenable.

SIGNIFICANCE OF MEASURING VARIATION/DISPERSION

The measure of dispersion is needed for the following purposes:

- It should be capable enough to be treated as a statistical technique.
- For determining the reliability of an average. It points out whether an average is representative of population or not.
- It provides basis for controlling the variables by determining the nature and cause of variability.
- Comparison of two or more series for their variability. Dispersion is a means for determining the uniformity or consistency.
- It facilitates the use of other measures of statistics.

METHODS OF STUDYING VARIATION/DISPERSION

The important methods of studying variation can be divided into:

1. Algebraic method:
   a. Methods of Limits:
      - Range
      - Interquartile Range and the Quartile Deviation
      - Percentile Range.
   b. Methods of Moments:
      - Mean Deviation
      - Standard Deviation
2. Graphical method: Algebraic Methods are again classified as:
   - Lorenz curve.

Alternate Classification

The methods of limits i.e. the range, quartile deviation and percentile range, are known as the **positional measures of variation** as it depends on the value at a particular position in the distribution. The mean deviation and the standard deviation are called calculation **measures of deviation** because all the values are needed for their calculation. The last Lorenz curve is a graphical method.

Measures of dispersion are of two types – **Absolute measure and Relative measure**. The **absolute measures of dispersion** are expressed in statistical units which are similar to the original data, for example, rupees, tones, etc. The absolute measure is not comparable, if the two data are expressed in two different units. For such cases we use the **relative measures of dispersion**. It is the ratio of absolute measure of dispersion to an appropriate average and also known as **coefficient of dispersion**. The term coefficient refers to a number which is independent of measurement of unit. Let us now study different methods of measuring variation/dispersion.

Algebraic Methods

RANGE

The simplest method of studying dispersion is Range. Range is defined as the difference between the largest value and the smallest value. It is calculated by the following formula:

Range = L – S

Where,

L = Largest value or the value of largest item, and

S = Smallest value or the value of smallest item.

The relative measure of dispersion also known as co-efficient of range is calculated by the following formula:

$$\text{Co-efficient of Range} = \frac{L - S}{L + S}$$

If the two distributions have same average, comparison of range indicates that the distribution with the smallest range will have less dispersion and the average is typical to the group.

Computation of Range

Individual Series

Illustration 1

The following are the incomes of 6 persons per day:

Persons	A	B	C	D	E	F
Income (Rs.)	200	210	280	160	250	350

Calculate the range and its co-efficient.

Solution

Range = L – S

L = 350 and S = 160

Range = 350 – 160 = 190

$$\text{Co-efficient of Range} = \frac{L-S}{L+S} = \frac{350-160}{350+160} = \frac{190}{510} = 0.372$$

Discrete Series

Illustration 2

From the following data, calculate Range and its Co-efficient:

Profit (Rs. in lakh)	5	10	15	20
No.of Companies	4	7	21	47

Solution

In the above data, L = 20 and S = 5

Range = L – S

Range = 20 – 5 = 15

$$\text{Co-efficient of Range} = \frac{L-S}{L+S} = \frac{20-5}{20+5} = \frac{15}{25} = 0.60$$

Illustration 3

Calculate the range and its co-efficient from the following data:

Wages per Week	0-10	10-20	20-30	30-40	40-50
No. of Workers	8	10	12	8	4

Solution

Range = L – S

L = 50 and S = 0

Range = 50 – 0 = 50

$$\text{Co-efficient of Range} = \frac{L-S}{L+S} = \frac{50-0}{50+0} = \frac{50}{50} = 1.00$$

Merits and Limitations

Merits:

i. Range is simple to understand and easy to compute.

ii. It gives quick picture of variability, as the time required to calculate range is minimum.

Limitations:

i. Range will not give the accurate picture of variability as it does not consider each and every item of the distribution.

ii. It fluctuates from sample to sample.

iii. The character of the distribution within the two extremes cannot be disclosed by the range.

Uses

Despite these limitations, the range is widely used in the following areas:

– For quality control.

– Useful in studying the fluctuations in the prices of stocks and shares.

– Useful in forecasting the weather.

– It is widely used in everyday life.

QUARTILE DEVIATION

Range, the measure of dispersion, which is based on the two extreme items of the distribution and fails to take into account the scatteredness within the range. To overcome this problem, a measure is developed known as Interquartile range. Interquartile range is a range that includes the middle 50 percent of the distribution i.e. it excludes the one quarter of the observation at the lower end and the other quarter at the upper end of the distribution. Interquartile range typically form what we call as distance measures of dispersion, the interquartile range looks at how far one should go from the median on either side of it before one-half of the data points are included. Quartiles are fractiles, which divide the data into four equal parts such that each quartile consists 25 percent of the data points. Quartiles are the highest data points in each of the first three points among the four parts. Interquartile range is the difference between the last value in the third quartile (usually denoted by Q_3) and the last value in the first quartile denoted by Q_1. In other words, the difference between the third quartile and first quartile is known as inter-quartile range. It is calculated by the following formula:

$$\text{Inter-quartile range} = Q_3 - Q_1$$

When the inter-quartile range is divided by 2, we get Quartile deviation or Semi-inter-quartile range. It is calculated by the following formula:

$$\text{QD} = \frac{Q_3 - Q_1}{2}$$

The average amount by which the first and third quartile differ from the median is given by quartile deviation. The exact 50 percent of the observations are not covered by Median $\pm$ QD in an asymmetrical distribution. Because in an asymmetrical distribution, the first and third quartiles are not equidistant from the median.

The absolute measure of dispersion is known as quartile deviation. The relative measure of dispersion is known as co-efficient of dispersion and is calculated by the following formula:

$$\text{Co-efficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Computation of Quartile Deviation

The procedure is the same as that of median discussed in the previous chapter. The only difference is – instead of taking

$$\text{Size of } \frac{n+1}{2} \text{ or } \frac{N+1}{2} \text{ or } \frac{N}{2}$$

When 'n' and 'N' are odd or even we have to take

$$\frac{n+1}{4} \text{ or } \frac{N+1}{4} \text{ or } \frac{N}{4}$$

depending upon the nature of the series.

Individual Series

The procedure is the same as that of median discussed in the previous chapter. The only difference is – instead of taking

$$\text{Size of } \frac{n+1}{2} \text{ or } \frac{N+1}{2}$$

When 'n' and 'N' are odd or even.

Illustration 4

From the following data, calculate the quartile deviation and its co-efficient:

Roll No.	1	2	3	4	5	6	7
Wages in Rs.	30	42	60	18	45	22	75

Solution

Calculation of Quartile Deviation

First arrange the data in ascending order:

18 22 30 42 45 60 75

Q_1= Size of $\frac{n+1}{4}$ th item = $\frac{7+1}{4} = 2\text{th item}$

Size of 2nd item is 22.

Q_3= Size of $\frac{3(n+1)}{4}$ th item = $\frac{3(7+1)}{4} = 6\text{th item}$

The Size of 6th item is 60. Thus, Q_3 is 60.

$$QD = \frac{Q_3 - Q_1}{2} = \frac{60-22}{2} = 19$$

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{60-22}{60+22} = \frac{38}{82} = 0.463$$

Discrete Series

Under discrete series, the computation of Q_1 and Q_3 is similar to individual series. However, the only difference is the value of N will be the sum of all frequencies (f) viz., no. of workers, total no. of companies etc.

Illustration 5

Compute the quartile deviation from the following data:

Marks	15	30	45	60	75	90
No. of Students	4	7	15	8	7	2

Solution

Calculation of Quartile Deviation

Marks	No. of Students		cf
15	4	0 + 4	4
30	7	4 + 7	11
45	15	4 + 7 + 15	26
60	8	4 + 7 + 15 + 8	34
75	7	4 + 7 + 15 + 8 + 7	41
90	2	4 + 7 + 15 + 8 + 7 + 2	43

Q_1 = Size of $\frac{N+1}{4}$ th item = $\frac{43+1}{4} = 11\text{th item}$

size marks of 11th item is 30. Thus, $Q_1 = 30$

Q_3 = Size of $\frac{3(N+1)}{4}$ th item = $\frac{3(43+1)}{4} = 33\text{th item}$

(there is 33rd item in cf, then we take the next immediate greater value i.e., 34th item in cf)

The size of 33rd item is 60. Thus, Q_3 is 60.

$$QD = \frac{Q_3 - Q_1}{2} = \frac{60-30}{2} = 15$$

$$\text{Co-efficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{60-30}{60+30} = \frac{30}{90} = 0.333$$

Continuous Series

For continuous series 'grouped data' the formula for calculating Q_1 and Q_3 is the same as median and is given by

$$Q_1 = L + \frac{N/4 - cf}{f} \times i \quad \text{and}$$

$$Q_3 = L + \left[\frac{3N/4 - cf}{f}\right] \text{ x } i$$

Where,

L = lower limit of the quartile class.
N = total frequency.
cf = cumulative frequency of class that is preceding to quartile class.
f = frequency of the quartile class.
i = width of the class interval.

The only difference between the two quartiles is that under Q_3 N is multiplied by 3 i.e., (3N) to get the position of the third quartile. It is to be noted that the value of N is similar to discrete series.

Illustration 6

Calculate the quartile deviation of the following distribution.

Percentage of Dividend Declared	No. of Companies
10-20	6
20-30	15
30-40	5
40-50	10
50-60	4
60-70	9
70-80	2
80-90	13

Solution

Class	No. of Companies (f)	Cumulative Frequency (cf)
10-20	6	6
20-30	15	21
30-40	5	26
40-50	10	36
50-60	4	40
60-70	9	49
70-80	2	51
80-90	13	64

Size of Q_1= Size of (N/4) = Size of (64/4) = Size of 16, which lies in 20-30 class.

$$Q_1 = L + \frac{N/4 - cf}{f} \times i$$

Thus, L = 20, cf = 6, f = 15 and i = 10

$$Q_1 = 20 + \left[\frac{64/4 - 6}{15}\right] \text{ x } 10 = 26.67$$

To determine the third quartile class, we have to identify the third quarter. We get this by finding the value of 3N/4. Where 3N/4 = 48. This lies in the class 60-70.

$$Q_3 = L + \left[\frac{3N/4 - cf}{f}\right] \text{ x } i$$

Here, L = 60, cf = 40, f = 9, i = 10

$$Q_3 = 60 + \left[\frac{(3 \times 64)/4 - 40}{9}\right] \times 10 = 68.89$$

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2} = \frac{68.89 - 26.67}{2}$$

$$= [68.89 - 26.67]/2 = 21.11.$$

$$\text{Co-efficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{68.89 - 26.67}{68.89 + 26.67} = \frac{42.22}{95.56} = 0.44$$

Merits and Limitations

Merits:

Quartile deviation is considered superior to range for following reasons:

- It is specially useful in the case of open and end distribution.
- As the quartile deviation is not affected by extreme values, it is useful in the case of erratic skewed distribution.

Limitations:

- Quartile deviation is not considered a good measure of dispersion, as it ignores the first and last 25% of the distribution.
- It cannot be manipulated mathematically.
- The values of quartile deviations are affected by sampling fluctuations.
- Quartile deviation is a positional average and does not show the scatteredness around the average but show the distance on the scale.

Because of above limitations quartile deviation is not used much in statistical inferences.

PERCENTILE RANGE

Box: 1 Percentile Range

Percentages are similar to Fractiles. If we say that in CFA level I, the pass percentage was 35, we understand that out of every 100 candidates who appeared for the examination, 35 candidates passed the examination. Similar to this, in a frequency distribution also there are always some elements, which lie at or below the given fractile. For instance, the median is 0.50 fractile as half of the data is less than or equal to this value. Further in any distribution, 25 percent of the data lies at or below the 0.25 fractile. Interfractile range is the difference between the values of two fractiles. Therefore, this is also necessarily a measure of dispersion (positional measure and involves algebraic method of calculation) between the two fractiles in a frequency distribution. Depending on the number of equal parts into which we divide the data we call them as deciles, percentiles and quartiles. A decile is a fractile that divides the data into ten equal parts, percentile is a fractile which divides the data into 100 equal parts and finally a quartile is the one which divides the data into four equal parts. While computing fractiles, we ought to arrange the elements in an increasing order. Percentile range of a set of data is calculated by the following formula:

Percentile Range = $P_{90} - P_{10}$

Where,

P_{90} is the 90th percentile and P_{10} is the 10th percentile. And the semi-percentile range will be $\frac{P_{90} - P_{10}}{2}$.

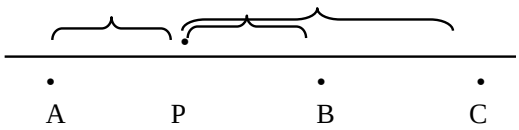
However, it is not used much in practice.

Methods of Moments

MEAN DEVIATION

The drawback of the range as a measure of dispersion is that it takes into account the values of only two data points – the largest and the smallest. The second measure i.e., quartile deviation ignores the first and last 25% of the distribution. So in the strict sense, it is not considered a measure of dispersion, and we should take the deviations from average in order to study the formation of distribution. This objective is achieved with the help of Mean deviation and standard deviation. One way to have a dispersion measure that takes into account, every data point is to use deviations. Deviation means difference. We can find the average of the deviations of the data points from a fixed value.

Figure



AP = Deviation of A from P.

BP = Deviation of B from P.

CP = Deviation of C from P.

The mean deviation is also known as Average deviation. Mean deviation is the average difference between the distribution items and the mean or median of that distribution. Theoretically 'the sum of deviations of items from median is minimum when signs are ignored'. However, in practice mean is used for calculating the mean deviation that's why the deviation is called Mean deviation.

Computation of Mean Deviation

Individual Series

$$MD = \frac{\sum |D|}{N}$$

Where,

$|D| = |X - A|$ |D| is read as mod D and it is an absolute value of the deviation which ignores plus and minus sign;

N = No. of observations, and

A = Mean or Median.

Steps:

1. Compute the mean or median of the series.
2. Take the deviation of item from the mean or median by ignoring the signs + (plus) and – (minus), and denote it as |D|.
3. Obtain the total of |D| i.e. $\sum |D|$.
4. Divide the sum obtained in the third step by the number of observations and thus we get the value of Mean Deviation.

Co-efficient of mean deviation is the relative measure of dispersion. It is obtained by dividing the mean deviation by its average (mean or median) used in the computation of mean deviation. The formula for calculating the co-efficient of Mean Deviation is

$$\text{Co-efficient of Mean Deviation} = \frac{\text{MD.}}{\text{Mean / Median}}$$

Illustration 7

Calculate the mean deviation and its co-efficient for the following data:

X	2	6	11	14	16	19	23

Solution

i. **Calculation of Mean Deviation from Median**

X	|D| = (X – Med)
2	12
6	8
11	3
14	0
16	2
19	5
23	9
N = 7	$\sum$ |D| = 39

Median = Size of $\frac{n+1}{2}$ th item = $\frac{7+1}{2} = 4$ th item

The value of 4^{th} item = 14 and therefore, median = 14.

$MD = \frac{\sum |D|}{n} = \frac{39}{7} = 5.57$

Co-efficient of Mean deviation = $\frac{MD}{Median} = \frac{5.57}{14} = 0.397$

ii. **Calculation of Mean Deviation from Mean**

X	|D| = (X – Mean)
2	11
6	7
11	2
14	1
16	3
19	6
23	10
	$\sum$ |D| = 40

$\bar{X} = \frac{\sum X}{n} = \frac{91}{7} = 13$

$MD = \frac{\sum |D|}{n} = \frac{40}{7} = 5.71$

Co-efficient of Mean deviation = $\frac{MD}{Mean} = \frac{5.71}{13} = 0.439$

Discrete Series

The formula for the computation of mean deviation in discrete series is given below:

$$\text{Mean Deviation} = \frac{\sum f\,|D|}{N}$$

Where,

|D| = deviations obtained from mean or median ignoring signs.

Steps:

i. Compute the mean or median of the series.

ii. Take the deviation of item from the mean or median by ignoring the signs (+) plus and (–) minus, and denote it as |D|.

iii. Multiply the deviations so obtained with their respective frequencies and obtain the total i.e. $\sum f|D|$.

iv. Divide the sum obtained in the third step by the number of observations and thus we get the value of Mean deviation.

Illustration 8

Calculate the Mean deviation from the following data:

Marks	5	10	15	20	25
No. of students	6	7	8	11	8

Solution

Calculation of Mean Deviation

Marks	No. of Students	cf	|D| = (X – Med)	f|D|
5	6	6	10	60
10	7	13	5	35
15	8	21	0	0
20	11	32	5	55
25	8	40	10	80
	N = 40			$\sum f\|D\|$ = 230

$$\text{Median} = \text{Size of } \frac{N+1}{2}\text{th item} = \frac{40+1}{2} = 20.5\,\text{th item}$$

The value of 20.5 item lies in marks = 15 and therefore, median = 15.

$$\text{MD} = \frac{\sum f\,|D|}{N} = \frac{230}{40} = 5.75$$

$$\text{Coefficient of Mean deviation} = \frac{\text{MD}}{\text{Median}} = \frac{5.75}{15} = 0.383$$

Illustration 9

Calculate the mean deviation from the mean for the following data:

Size	4	6	8	10	12	14	16
Frequency	2	1	3	6	4	3	1

Solution

Calculation of Mean Deviation and its Coefficient

Size	Frequency	fx	|D| = (X – Mean)	f|D|
4	2	8	6.2	12.4
6	1	6	4.2	4.2
8	3	24	2.2	6.6
10	6	60	0.2	1.2
12	4	48	1.8	7.2
14	3	42	3.8	11.4
16	1	16	5.8	5.8
	N = 20	$\sum$ fx = 204		$\sum$ f|D| = 48.8

$$\bar{X} = \frac{\sum fX}{N} = \frac{204}{20} = 10.2$$

$$MD = \frac{\sum f\,|D|}{N} = \frac{48.8}{20} = 2.44$$

$$\text{Coefficient of Mean deviation} = \frac{MD}{Mean} = \frac{2.44}{10.2} = 0.239$$

Continuous Series

The procedure for calculating the mean deviation in continuous series is similar to discrete series. However, in continuous series we have to calculate the mid-points of the various classes and then we take deviations of these mid-points from the mean or the median. The formula of mean deviation in continuous series is given below:

$$\text{Mean Deviation} = \frac{\sum f\,|D|}{N}$$

Illustration 10

Calculate the Mean Deviation from Mean and Median for the following data:

Class Interval	2-4	4-6	6-8	8-10
Frequency	3	4	2	1

Solution

Calculation of Mean Deviation and its Coefficient

CI	f	m (mid points)	fm	|D| = (X – Mean)	f|D|
2-4	3	3	9	2.2	6.6
4-6	4	5	20	0.2	0.8
6-8	2	7	14	1.8	3.6
8-10	1	9	9	3.8	3.8
	N = 10		$\sum$ fx = 52		$\sum$ f|D| = 14.8

$$\bar{X} = \frac{\sum fx}{N} = \frac{52}{10} = 5.2$$

$$MD = \frac{\sum f\,|D|}{N} = \frac{14.8}{10} = 1.48$$

$$\text{Co-efficient of Mean Deviation} = \frac{MD}{Mean} = \frac{1.48}{5.2} = 0.285$$

Calculation of Mean Deviation

CI	Frequency	cf	m	|D| = (X – Med)	f|D|
2-4	3	3	3	2	6
4-6	4	7	5	0	0
6-8	2	9	7	2	4
8-10	1	10	9	4	4
	N = 10				$\sum$ f|D| = 14

Median = Size of $\frac{N}{2}$ th item = $\frac{10}{2} = 5$ th item

The value of 5th item lies in class 4-6 and therefore,

$$\text{Median} = L + \frac{N/2 - Cf}{f} \times i = 4 + \frac{5-3}{4} \times 2 = 4 + 1 = 5$$

$$MD = \frac{\sum f|D|}{N} = \frac{14}{10} = 1.40$$

$$\text{Coefficient of Mean deviation} = \frac{MD}{\text{Median}} = \frac{1.40}{5} = 0.28$$

Merits and Limitations

Merits:

i. It is simple to understand and easy to compute.

ii. It is a better measure of dispersion when compared to range and quartile deviation because it includes all observations in its calculation.

iii. Mean deviation is less affected by the extreme values when compared to standard deviation.

iv. Mean deviation is considered a true and accurate measure of dispersion.

Limitations:

i. The basic and strong objection against mean deviation is that it considers the absolute value of the deviations and ignores the algebraic signs, which are mathematically unsound and illogical.

ii. Further mathematical treatment is not possible due to first limitation.

iii. It is rarely used in sociological sciences.

iv. It does not give accurate result because the deviations are taken from median which will not give satisfactory result in case of high degree of variability.

Despite the above limitations, the mean deviation has wider practical utility in Economics and Business statistics. It is popular because of its simplicity, accuracy in understanding and computations.

STANDARD DEVIATION AND CO-EFFICIENT OF VARIATION

Standard Deviation

The standard deviation is an improved measure of dispersion. This concept was introduced by Karl Pearson in 1893 and denoted by a small Greek alphabet σ it is also known as Root Mean Square deviation because it is the square root of the arithmetic mean of the squared deviation from their arithmetic mean. It is an absolute measure of dispersion.

Calculation of the standard deviation involves the following four steps:

i. Calculation of deviations of the observations from the mean.

ii. Squaring each deviation.

iii. Finding the mean of the squared deviations obtained in step (ii).

iv. Taking the positive square root of the mean found in step (iii).

Individual Series

a. **When the deviations are taken from actual mean:** When the deviations are taken from actual mean, we apply the following formula for calculating the standard deviation:

$$\sigma = \sqrt{\frac{1}{n}\sum (X - \bar{X})^2} = \sqrt{\frac{\sum x^2}{n}}$$

Where, $x = (X - \bar{X})$

Or

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \left[\frac{\sum X}{N}\right]^2}$$

Illustration 11

Consider the observations 2, 6 and 7 and compute the standard deviation of these observations:

Solution

Calculation of Standard Deviation

	Observation X	(Step 1) $X - \bar{X}$	(Steps 2 & 3) $(X - \bar{X})^2$	X^2
	2	–3	9	4
	6	1	1	36
	7	2	4	49
Total	15	0	14	$\sum X^2 = 89$

Mean $\bar{X} = \frac{15}{3} = 5$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} = \frac{14}{3}$$

$= \sqrt{4.667} = 2.16$ is the standard deviation.

Or $$\sigma = \sqrt{\frac{\sum X^2}{N} - \left[\frac{\sum X}{N}\right]^2}$$

$$= \sqrt{\frac{89}{3} - [5]^2}$$

$$= \sqrt{29.67 - 25} = \sqrt{4.67} = 2.16$$

b. **When the deviations are taken from the assumed mean:** When the deviations taken from the actual mean are tedious and cumbersome, in such case we obtain the deviations from the assumed mean. The following formula is used:

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left[\frac{\sum d}{N}\right]^2}$$

Where, d = (X – A), and A = assumed mean.

Steps: When the deviations are taken from the assumed mean, the following steps are required:

i. Assume an item and take the deviation of the item from the assumed mean, denote it by d and obtain its total i.e. $\sum d$.

ii. Square the deviations and obtain its total i.e. $\sum d^2$.

iii. Substitute all the above values in the formula and obtain the value of σ.

Illustration 12

Calculate the standard deviation for the following data:

Size	4.5	14.5	24.5	34.5	44.5	54.5	64.5

Solution

Calculation of Standard Deviation

X	(X – 34.5) = d	d^2
4.5	–30	900
14.5	–20	400
24.5	–10	100
34.5	0	0
44.5	+10	100
54.5	+20	400
64.5	+30	900
N = 7	$\sum d = 0$	$\sum d^2 = 2800$

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left[\frac{\sum d}{N}\right]^2} = \sqrt{\frac{2800}{7} - \left[\frac{0}{7}\right]^2} = \sqrt{400}$$

$\sigma = 20$

Discrete Series

a. **When the Deviations are taken from the Actual Mean**

The formula applied is $\sigma = \sqrt{\frac{\sum f x^2}{N}}$

Where, N = frequency of the variable, x = $(X - \bar{X})$

Illustration 13

Calculate the standard deviation from the data given below:

Size of item	4	5	6	7	8	9	10
Frequency	6	8	10	15	12	10	9

Solution

Calculation of Standard Deviation

Size of Item (X)	Frequency (f)	fX	$x = (X - \bar{X})$	x^2	fx^2
4	6	24	–3.21	10.3041	61.824
5	8	40	–2.21	4.8841	39.072
6	10	60	–1.21	1.4641	14.641
7	15	105	–0.21	.0441	0.661
8	12	96	+0.79	.6241	7.489
9	10	90	+1.79	3.2041	32.041
10	9	90	+2.79	7.7841	70.057
	N = 70	505			$\sum fx^2$ = 225.785

$$\bar{x} = \frac{505}{70} = 7.21$$

$$\sigma = \sqrt{\frac{\sum f x^2}{N}} = \sqrt{\frac{225.785}{70}} = \sqrt{3.225} = 1.796$$

b. **When the Deviations are taken from the Assumed Mean**

When we take the deviations from the assumed mean, following formula is used:

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N}\right]^2}$$

Steps: When the deviations are taken from the assumed mean, following steps are followed:

i. Assume an item and take the deviation of the item from the assumed mean and denote it by d.

ii. Multiply the deviations with their respective frequency and obtain its total i.e. $\sum fd$.

iii. Square the deviations, multiply with their respective frequency and obtain its total i.e. $\sum fd^2$.

iv. Substitute all the above values in the formula and obtain the value of σ.

Illustration 14

Calculate the standard deviation from the data given below:

Size of item	4	5	6	7	8	9	10
Frequency	6	8	10	15	12	10	9

Solution

Calculation of Standard deviation

Size of Item (X)	Frequency (f)	(X – 7) = d	fd	d^2	fd^2
4	6	–3	–18	9	54
5	8	–2	–16	4	32
6	10	–1	–10	1	10
7	15	0	0	0	0
8	12	+1	+12	1	12
9	10	+2	+20	4	40
10	9	+3	+27	9	81
	N = 70		$\sum fd$ =+15		$\sum fd^2$ = 229

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N}\right]^2} = \sqrt{\frac{229}{70} - \left[\frac{15}{70}\right]^2} = \sqrt{3.271 - .0459} = \sqrt{3.225}$$

$$\sigma = 1.796$$

Continuous Series

In continuous series, the methods discussed under the discrete series can be used. Apart from the two methods, the third method that can be used is step deviation method:

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$

Where,

d = (m – A)/i.

i = Class Interval.

A = Assumed Mean.

m = Mid-point of the class interval.

The steps involved are:

- Find mean for grouped data.
- Find deviations from mean for grouped data.
- Find squares of the above deviations.
- Total the squared deviations taking frequency into account.
- Calculate square root.

Illustration 15

A security analyst studied hundred companies and obtained the following Return on Investment (ROI) data for the year 2000.

Returns %	0-10	10-20	20-30	30-40
No. of companies	19	32	41	8

We can find how the ROI of the company varies with the mean ROI by calculating the standard deviations for the above data.

Solution

Calculation of Standard Deviation

Return on Investment	Mid-point	No. of Companies	Deviation		
%	m	f	fX	$X - \bar{X}$	$f(X - \bar{X})^2$
0-10	5	19	95	–13.8	3618.36
10-20	15	32	480	–3.8	462.08
20-30	25	41	1025	6.2	1576.04
30-40	35	8	280	16.2	2099.52
Total		100	1880		7756.00

Mean $\bar{X} = \frac{\sum fX}{N} = \frac{1880}{100} = 18.8\%$

Standard Deviation = $\sqrt{\frac{\sum fx^2}{N}} = \sqrt{\frac{7756}{100}} = 8.81\%$

Thus, the standard deviation for the return on investment is 8.8%.

In such a calculation, we always assume that all the observations in a class interval are located at the mid-point of the class. For example, the first class interval has mid-point 5 and frequency 19. Hence, the assumption is that all the 19 companies have an ROI of exactly 5%.

Alternative Method

In practice, it is the step deviation method that is most used for calculating standard deviation from grouped data.

When the step deviation method is used, deviations of mid-points from an assumed mean are taken and divided by the width of the class interval, i.e. 'i'. The formula applied is

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$

Where,

d = (x – A)/i.

i = Common Factor that is used.

A = Assumed Mean.

x = Mid-point of the class interval.

Take A = 15. Here i = 10

Dividend Declared	Mid-point (x)	No. of Companies (f)	(x – 15)/10 (d)	fd	fd^2
0-10	5	19	–1	–19	19
10-20	15	32	0	0	0
20-30	25	41	1	41	41
30-40	35	8	2	16	32
Total		100		38	92

$$\sigma = \sqrt{\frac{92}{100} - \left(\frac{38}{100}\right)^2} \times 10 = 0.8806 \times 10 = 8.806 \text{ or } 8.81$$

Illustration 16

The following are the sales figures of some Granite Industries for the year 2000-2001. You are required to calculate standard deviation.

Company	Sales (Rs. in lakh)
Ankit Granites	1.86
Deccan Granites	4.54
BTW Granites	30.93
Pacific Granites	12.80
Jaswal Granites	3.45
Grapco Granites	4.42

Solution

Calculation of Standard Deviation

Industry Name	Sales X	$X - \bar{X}$	$(X - \bar{X})^2$
Ankit Granites	1.86	–7.81	61.00
Deccan Granites	4.54	–5.13	26.32
BTW Granites	30.93	21.26	451.99
Pacific Granites	12.80	3.13	9.80
Jaswal Granites	3.45	–6.22	38.69
Grapco Granites	4.42	–5.25	27.55
Total	58.00		615.35

$\bar{X} = (\sum X)/n = 58.00/6 = 9.67$

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}} = \sqrt{\frac{615.35}{5}} = 11.09$$

The individual sales of the granite industries deviate to the extent of Rs.11.09 lakh from the average sales of the industry.

Properties of Standard Deviation

i. The value of standard deviation remains the same if, in a series each of the observation is increased or decreased by a constant quantity. In statistical language, we say standard deviation is independent of change of origin.

For example, for the observations 3, 10 and 12, then $\bar{X} = 8.33$ and $\sigma = 3.859$.

If we increase the value of each observation by 4.5, we get the observations 7.5, 14.5 and 16.5.

Now, $\bar{X} = 12.833$ and $\sigma = 3.859$

Hence, although mean has increased by 4.5, σ remains the same.

ii. For a given series, if each observation is multiplied or divided by a constant quantity, standard deviation will also be similarly affected.

Consider the observations 3, 10, 12. $\sigma = 3.859$ as shown in the above calculation.

Suppose we multiply each observation by 6, the observations become 18, 60 and 72.

$\bar{X} = 50$

$$\sigma = \sqrt{\frac{(18-50)^2 + (60-50)^2 + (72-50)^2}{3}} = 23.152$$

Which is nothing but the earlier σ multiplied by 6, i.e., 3.859 x 6.

In short, standard deviation is independent of any change of origin, but dependent on the change of scale.

iii. Standard deviation is the minimum root-mean-square deviation. In other words, the sum of the squares of the deviations of items of any series from a value other than the arithmetic mean would always be greater.

iv. Just as it is possible to compute combined mean of two or more groups, it is also possible to compute combined standard deviation of two or more groups. Combined standard deviation denoted by σ_{12} is computed as follows:

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

where,

$\bar{X}_1$ = Mean of first group.

$\bar{X}_2$ = Mean of second group.

σ_1 = Standard deviation of first group.

σ_2 = Standard deviation of second group.

N_1 = Number of observations in the first group.

N_2 = Number of observations in the second group.

$d_1 = \bar{X}_1 - \bar{X}$

$d_2 = \bar{X}_2 - \bar{X}$

$\bar{X} = (N_1 \bar{X}_1 + N_2 \bar{X}_2) / (N_1 + N_2)$

Illustration 17

Consider the following 2 sets of observations:

Set 1	3	10	12
Set 2	6	4	1,9

For set 1, $\bar{X}_1 = 8.333$ and $\sigma_1 = 3.859$

For set 2, $\bar{X}_2 = 5$ and $\sigma_2 = 2.915$

(Consider these 2 sets of observations as populations).

Combined mean $\bar{X}$

$$= \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2} = \frac{(3 \times 8.333) + (4 \times 5)}{3 + 4} = 6.429$$

$d_1 = \bar{X}_1 - \bar{X} = 8.333 - 6.429 = 1.904$

$d_2 = \bar{X}_2 - \bar{X} = 5 - 6.429 = -1.429$

σ [i.e. standard deviation for the combined set of 7 observations]

$$= \sqrt{\frac{(3 \times 3.859^2) + (4 \times 2.915^2) + (3 \times 1.904^2) + [4 \times (-1.429)^2]}{3 + 4}}$$

$$= 3.736$$

We can verify this by calculating σ using the 7 observations which verifies the rule.

$$\sigma^2 = \frac{\begin{pmatrix}(3 - 6.249)^2 + (10 - 6.429)^2 + (12 - 6.429)^2 \\ (3 - 6.249)^2 + (10 - 6.429)^2 + (12 - 6.429)^2\end{pmatrix}}{7} = 13.96$$

$\sigma = \sqrt{13.96} = 3.736$ which verifies the rule.

In a later chapter, we will present another method of computing the standard deviation of a portfolio using covariances.

Variance

The term variance was used to describe the square of the standard deviation by R.A.Fisher. The concept of variance is highly important in areas where it is possible to split the total into several parts, each attributable to one of the factors causing variation in their original series. Variance is denoted by σ^2 in the case of population and s^2 in the case of sample.

Variance is the average squared deviation from the arithmetic mean. The smaller the value of σ^2 ,the lesser the variability or greater the uniformity in the population.

Illustration 18

The data given below relates to the return in terms of percentage on the equity stocks of TS Co. and CI Ltd. for the past six years by the use of which the mean return and the standard deviation for the individual securities can be calculated.

Year	TS Co. %	CI (%)
1996-1997	35	30
1997-1998	30	35
1998-1999	30	35
1999-2000	30	25
2000-2001	25	20
2001-2002	25	30

Solution

Mean Return	29.17%	29.17%
Standard Deviation	3.76%	5.85%

Now, even though both the stocks yield the same percentage of return of 29.17%, an investor would opt for investing in the scrips of TS Co. which has smaller standard deviation. The desire for more uniformity (on small variation or great consistency) in the return leads the investor to place greater confidence in investing in TS Co. scrips, whose return shows a smaller standard deviation.

Merits and Limitations

Merits:

i. It is defined rigidly and therefore, it is the dependable measure of dispersion.

ii. It is capable of algebraic treatmentie., it can be used to test the consistency of the data.

iii. Unlike Mean Deviation, Combined Standard deviation for given two or more series can be computed if $\overline{X}$s and SDs are given.

iv. It is based on all the terms or observations. Hence, it is more reliable.

v. It is the best measure because it signifies that sum of squares of deviations from $\bar{X}$ is minimum.

vi. It is the suitable measure for Normal Distribution.

vii. It is basically based on arithmetic mean. Hence, it possesses many fine qualities of arithmetic mean.

Demerits:

i. It involves difficulty in computation.

ii. It takes the support of co-efficient of variation to make comparison of two series.

iii. The extreme terms make the impact too much, therefore in some cases co-efficient of quartile deviation or co-efficient of mean deviation is regarded best measure when compared to it. If extreme items differ largely, they make a heavy change when deviations are squared.

Coefficient of Variation

The standard deviation discussed above is an absolute measure of dispersion. The corresponding relative measure is known as the coefficient of variation. This relative measure of dispersion based upon standard deviation is also called coefficient of standard deviation.

$$\text{Coefficient of Variation} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

It is used in such problems where we want to compare the variability, homogeneity, stability, uniformity and consistency of two or more series. That series for which the coefficient of variation is greater is said to be more variable or conversely less consistent, less uniform, less stable or less homogeneous. On the other hand, the series for which the Coefficient of Variation is less, is said to be less variable or more consistent, more uniform, more stable or more homogeneous.

Illustration 19

The following data pertains to the consumption of rice during a year by various families:

Consumption of Rice in kgs.	No. of Families
0-50	20
50-100	24
100-150	32
150-200	36
200-250	28
250-300	20
300-350	16

You are required to calculate coefficient of variation based on the above data.

Solution

Consumption of Rice (in kgs.)	No of families (f)	Mid-Value (x)	$d' = \frac{X - 175}{50}$	d'^2	fd'	fd'^2
0-50	20	25	3	9	–60	180
50-100	24	75	–2	4	–48	96
100-150	32	125	–1	1	–32	32
150-200	36	175	0	0	0	0
200-250	28	225	1	1	28	28
250-300	20	275	2	4	40	80
300-350	16	325	3	9	48	144
	$\sum f = N(n)=176$				$\sum fd' = -24$	$\sum fd'^2 = 560$

Arithmetic mean, $\overline{X} = A + \frac{\sum fd'}{N} \times C$

$A = 175,\ \sum fd' = 24,\ \sum fd'^2 = 560,\ N=176,\ C = 50$

$$= 175 + \frac{(-24)}{176} \times 50 = 175 - 6.818 = 168.182$$

$$\text{Standard deviation} = \sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2} \times C$$

$$= \sqrt{\frac{560}{176} - \left(\frac{-24}{176}\right)^2} \times 50$$

$$= \sqrt{3.182 - 0.0186}\ x\ 50$$

$$= \sqrt{3.1634}\ \text{x}\ 50 = 88.93 \text{ (approx.)}$$

$$\text{Co-efficient of variation} = \frac{\sigma}{\overline{X}} \times 100 = \frac{88.93}{168.182} \times 100 = 52.88\% \text{ (approx.)}$$

Obtaining Correct Standard Deviation when Incorrect Values are Given

We know that

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \left[\frac{\sum X}{N}\right]^2}$$

To correct the incorrect value of Standard Deviation (SD) when some wrong items are included in the series, we have to first obtain correct $\overline{X}$.

Total of terms = $\sum X$

Corrected $\sum X$ = Incorrect $\sum X$ – wrong values + correct values

$$\text{Corrected } \overline{X} = \frac{\text{Corrected} \sum X}{n}$$

We can also find incorrect $\sum X^2$ by substituting the given data in the following way:

Corrected $\sum X^2$= Incorrect $\sum X^2$ – sum of squares of wrong values

+ sum of squares of correct values

Substituting these values in given formula, we get

$$\text{Corrected SD} = \sigma = \sqrt{\frac{\text{Corrected} \sum X^2}{n} - \left[\frac{\text{Corrected} \sum X}{n}\right]^2}$$

Illustration 20

The Mean and Standard Deviation(SD) of a series where N=100 is 20 and 4 respectively. However, later on it was found that a term 20 was misread as 22. Find out the correct mean and correct SD.

Solution

We are given N = 100, $\bar{X}$ = 20

Since $\bar{X} = \frac{\sum x}{N}$

$\sum x = N\ \bar{X}$ = 100 x 20 = 2000

But this is not correct $\sum x$

Correct $\sum x$ = Incorrect $\sum x$ – wrong item + correct item

= 2000 – 22 + 20 = 1998

$\therefore$ Correct $\bar{X} = \frac{\text{Corrected} \sum x}{N}$

$= \frac{1998}{100} = 19.98$

$SD = \sigma = \sqrt{\frac{\sum X^2}{N} - \left[\frac{\sum X}{N}\right]^2}$

$= 4 = \sqrt{\frac{\sum X^2}{100} - [20]^2}$

Squaring on both sides we get

$16 = \frac{\sum X^2}{100} - 400$

$= \frac{\sum X^2}{100} = 400 + 16 = 416$

$= \frac{\sum X^2}{100} = 416 \times 100 = 41,600$

But this is not correct

Correct SD = $\sum X^2 = 41,600 - (22)^2 + (20)^2$

= 41,600 – 484 +400 = 41,516

Correct SD = $\sqrt{\frac{41516}{100} - (19.98)^2}$

$= \sqrt{415.16 - 399.2}$

$= \sqrt{15.96} = 3.99$

RELATIONSHIP BETWEEN QUARTILE DEVIATION (QD), STANDARD DEVIATION (SD) AND MEAN DEVIATION (MD)

QD = $\frac{2}{3}$ SD and MD = $\frac{4}{5}$ SD

Or (in terms of SD)

SD = $\frac{3}{2}$ QD and SD = $\frac{5}{4}$ MD

Aternatively,

$\overline{X} \pm QD$ covers 50% of the total terms.

$\overline{X} \pm MD$ covers 50% of the total terms

$\overline{X} \pm SD$ covers 50% of the total terms.

$\overline{X} \pm QD$ covers 50% of the total terms.

$\overline{X} \pm MD$ covers 57.51% of the total terms.

$\overline{X} \pm SD$ covers 68.27% of the total terms.

GRAPHICAL METHOD – LORENZ CURVE

The graphical method of studying the dispersion of the distribution is known as Lorenz Curve. Max. O. Lorenz, an economic statistician first developed it for measuring the economic inequalities in the distribution of income and wealth between the different countries or between the different period of time. Today Lorenz curve is used in business for studying the inequalities or disparities in the distribution of wages, profits, turnover and production. A unique feature of Lorenz curve is that it deals with cumulative values of variable and frequencies rather than the absolute value and given frequencies. Lorenz Curve is the cumulative percentage curve consisting of percentage of items combined with percentage of other things. Lorenz Curve is simple and consists of following steps:

i. Both the size of the item (variable values) and the frequencies are cumulated. The grand total for each is taken as 100 and the cumulated total for the variable and the frequencies are expressed as percentage of the corresponding grand total.

ii. Start the X-axis from 0 to 100 and take the percentage of cumulated frequencies.

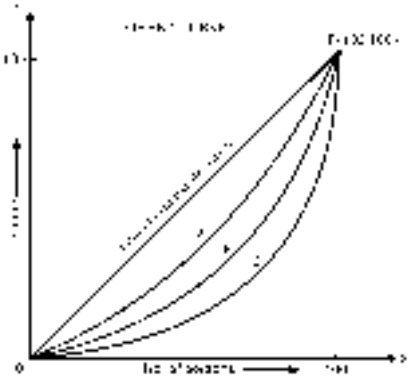
iii. Start the Y-axis from 0 to 100 and take the percentage of cumulated values of variables.

iv. Draw a diagonal line y = x, joining O the origin (0,0) with the point P(100,100). The line OP is called a line of equal distribution and make an angle of 45 degrees. We get the same percent on X as on Y for any point on this diagonal.

v. Plot the (Y) i.e. the percentages of the cumulated values of the variables against the (X) i.e. the percentages of the corresponding cumulated frequencies for the given distribution and join these points with a smooth free hand curve. For any given distribution this will never cross the line of equal distribution OP. It will always lie below OP unless the distribution is uniform in which case it will coincide with OP. The greater the variability, the greater is the distance of the curve from OP.

In the diagram given below, OP is the line of equal distribution. The points on the OAP indicates a less degree of variability when compared to the points on the OBP curve. The variability is still greater for the points lying on the curve OCP.

Figure



ADDITIONAL ILLUSTRATIONS

Illustration 1

Consider the following data: 80 , 77, 78, 76, 77, 74, 79, 77, 74, 75, 78.

Calculate the Standard deviation and Coefficient of variation.

Solution

Calculation of Standard Deviation

X	d = (X – 77)	d^2
80	+3	9
77	0	0
78	+1	1
76	–1	1
77	0	0
74	–3	9
79	+2	4
77	0	0
75	–2	4
78	+1	1
N = 10	$\sum d = 0$	$\sum d^2 = 29$

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left[\frac{\sum d}{N}\right]^2}$$

$$= \sqrt{\frac{29}{10} - \left[\frac{0}{10}\right]^2}$$

$$= \sqrt{2.9 - 0} = 1.703$$

Coefficient of Variation:

$$CV = \frac{\sigma}{\bar{X}} \times 100$$

$$= \frac{1.703}{77} \times 100 = 2.212$$

$$\bar{X} = A + \frac{\sum d}{N} = 77 + \frac{0}{10} = 77$$

Illustration 2

Calculate the Mean and standard deviation and coefficient of variation for the data given below:

Class-interval	Frequency
93-97	2
98-102	5
103-107	12
108-112	17
113-117	14
118-122	6
123-127	3
128-132	1

Solution

Calculation of Mean and Standard Deviation

Class-interval	Frequency	m	(X – 110)=d	d^2	fd	fd^2
93-97	2	95	–15	225	–30	450
98-102	5	100	–10	100	–50	500
103-107	12	105	–5	25	–60	300
108-112	17	110	0	0	0	0
113-117	14	115	5	25	70	350
118-122	6	120	10	100	60	600
123-127	3	125	15	225	45	675
128-132	1	130	20	400	20	400
	N = 60				55	3275

$$\bar{X} = A + \frac{\sum fd}{N} = 110 + \frac{55}{60} = 110.9167$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N}\right]^2}$$

$$= \sqrt{\frac{3275}{60} - \left[\frac{55}{60}\right]^2}$$

$$= \sqrt{53.74306} = 7.331$$

$$CV = \frac{\sigma}{\bar{X}} \times 100 = \frac{7.331}{110.9167} \times 100 = 6.6095\%$$

Illustration 3

For the data given below, calculate the:

a. Range.

b. Interfractile range between the fourth and sixth deciles.

98	69	58	87	73	89	83	65	82	63
88	91	77	68	94	86	96	89	98	85
55	59	87	84	59	82	73	95	68	81

Solution

a. Range = L – S = 98 – 55 = 43

b. The 4^{th} decile = 77, and the 6^{th} decile = 85. The interfractile range between 4^{th} and 6^{th} decile = 85 – 77 = 8.

Illustration 4

Calculate Median, Quartile deviation and its coefficient from the following frequency distribution:

Families	0	1	2	3	4	5	6
No. of Children	7	10	16	25	18	11	8

Solution

Calculation of Quartile Deviation

Families (x)	No. of Children (f)	cf
0	7	7
1	10	17
2	16	33
3	25	58
4	18	76
5	11	87
6	8	95

Median = Size of $\frac{N+1}{2}$ th item = $\frac{95+1}{2} = 48\text{th item}$

The size of 48^{th} item is 3. Therefore, Median = 3.

Q_1 = Size of $\frac{N+1}{4}$ th item = $\frac{95+1}{4} = 24\text{th item}$

The size of 24^{th} item is 2. Hence $Q_1 = 2$.

Q_3 = Size of $\frac{3(N+1)}{4}$ th item = $\frac{3(95+1)}{4} = 72\text{th item}$

Hence, Q_3 is 4.

$$QD = \frac{Q_3 - Q_1}{2} = \frac{4-2}{2} = 2$$

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{4-2}{4+2} = \frac{2}{6} = 0.33$$

Illustration 5

The causal life insurance company is considering purchase of new fleet of company cars. The financial department's director Mr. Mehra sampled 39 employees to determine the number of miles each drove over 1-year period. The results of the study are given below. Calculate the quartile deviation and interquartile range.

3600	4200	4700	4900	5300	5700	6700	7300
7700	8100	8300	8400	8700	8700	8900	9300
9500	9500	9700	10000	10300	10500	10700	10800
11000	11300	11300	11800	12100	12700	12900	13100
13500	13800	14600	14900	16300	17200	18500	

Solution

$$Q_1 = \frac{N+1}{4} = \quad Q_1 = \frac{40}{4} = 10\text{th item}$$

The value of the 10^{th} item = 8,100 and the Q_1 = 8,100

$$Q_3 = \frac{3(N+1)}{4} = \frac{3\times 40}{4} = 30\text{th item}$$

The value of 30^{th} item =12,700 and the Q_3 = 12,700

$$QD = \frac{Q_3 - Q_1}{2}$$

$$= \frac{12,700 - 8,100}{2} = \frac{4,600}{2} = 2,300 \text{ miles}$$

Interquartile Range $= Q_3 - Q_1 = 12,700 - 8100 = 4,600$

Illustration 6

The following table gives the distribution of monthly incomes of 500 workers in a factory. Calculate Quartile deviation and its coefficient.

Monthly Income (Rs.)	No. of Workers
Below Rs.100	10
100-150	25
150-200	145
200-250	220
250-300	70
300 and above	30

Solution

Calculation of Quartile Deviation and its Coefficient

Monthly income (Rs.)	No. of Workers (f)	cf
Below-100	10	10
100-150	25	35
150-200	145	180
200-250	220	400
250-300	70	470
300 and above	30	500

Q_1 = Size of $\frac{N}{4}$ th item = $\frac{500}{4} = 125$th item

$$Q_1 = L + \frac{N/4 - cf}{f} \times i$$

$$= 150 + \frac{125 - 35}{145} \times 50$$

$$= 150 + \frac{90}{145} \times 50$$

$$= 150 + 31.03 = 181.03$$

Q_3 = Size of $\frac{3N}{4}$ th item = $\frac{3(500)}{4} = 375$th item

$$Q_3 = L + \frac{3N/4 - cf}{f} \times i$$

$$= 200 + \frac{375 - 180}{220} \times 50$$

$$= 200 + \frac{195}{220} \times 50$$

$$= 200 + 44.32 = 244.32$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{244.32 - 181.03}{2} = 31.645$$

Coefficient of Quartile Deviation

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{244.32 - 181.03}{244.32 + 181.03} = \frac{63.29}{425.35} = +0.148$$

Illustration 7

Calculate the mean deviation and its coefficient from the income group of 5 persons.

Rs.	3000	3200	3400	3600	3800

Solution

Calculation of Mean Deviation

Income (Rs.)	Deviations from Median |D|
3,000	400
3,200	200
3,400	0
3,600	200
3,800	400
	$\sum \|D\| = 1200$

Median = Size of $\frac{N+1}{2}$ th item = $\frac{5+1}{2} = 3^{rd}$ item.

The size of third item = 3,400.

$$MD = \frac{\sum |D|}{N} = \frac{1200}{5} = 240$$

$$\text{Coefficient of MD} = \frac{MD}{\text{Median}} = \frac{240}{3,400} = 0.07$$

Illustration 8

Confederate Stereos, a wholesaler, was contemplating in becoming the supplier to three retailers, but inventory shortage has forced Confederate to select only one. Confederate's credit manager is evaluating the credit record of these three retailers. Over the past 5 years, these retailer's accounts receivable have been outstanding for the following average number of days. The credit manager feels that consistency, in addition to lowest average, is important. Based on relative dispersion, which retailer would make the best customer?

Suresh	Raman	Vijay
62.1	62.5	61.9
61.8	61.9	61.9
63.2	62.8	62.9
62.9	63.0	63.7
61.7	60.7	61.5

Solution

Suresh

Accounts Receivable (X)	$(X - \bar{X})$	$(X - \bar{X})^2$
62.1	–0.24	0.0576
61.8	–0.54	0.2916
63.2	+0.86	0.7396
62.9	+0.56	0.3136
61.7	–0.64	0.4096
311.7		1.812

Mean = 311.7/5 = 62.34

Variance = 1.812/(5 – 1) = 0.453

SD = $\sqrt{(0.453)}$ = 0.6731

CV = [(0.6731)/(62.34)] x 100 = 1.0796%

Raman

Accounts Receivable (X)	$(X - \bar{X})$	$(X - \bar{X})^2$
62.5	+0.32	0.1024
61.9	–0.28	0.0784
62.8	+0.62	0.3844
63.0	+0.82	0.6724
60.7	–1.48	2.1904
310.9		3.428

Mean = 310.9/5 = 62.18

Variance = 3.428/(5 – 1) = 0.857

SD = $\sqrt{(0.857)}$ = 0.9257

CV = [(0.9257)/(62.18)] x 100 =1.4888%

Vijay

Accounts Receivable (X)	$(X - \bar{X})$	$(X - \bar{X})^2$
61.9	–0.48	0.2304
61.9	–0.48	0.2304
62.9	+0.52	0.2704
63.7	+1.32	1.7424
61.5	–0.88	0.7744
311.9		3.248

Mean = 311.9/5 = 62.38

Variance = 3.248/(5 – 1) = 0.812

SD = $\sqrt{(0.812)}$ = 0.9011

CV = [(0.9011)/(62.38)] x 100 = 1.4445%

Comparing the CVs, Suresh found more consistent and he would make the best customer.

Illustration 9

The mean and standard deviation of a sample consisting of 50 observations were worked out to be 40 and 22 respectively. Subsequently, it was found out that there was a mistake in the process and one of the observations included in the sample was 150 instead of 50. Calculate the correct mean and correct Standard deviation.

Solution

Incorrect sample:

$\bar{x} = \frac{\sum x}{n}$ or $n\bar{x} = \sum x = 50 \times 40 = 2000$

Variance, $\sigma^2 = 22^2 = 484$

$$\sigma^2 = \frac{\sum x^2}{n-1} - \frac{n\bar{x}^2}{n-1}$$

$\therefore 484 = \frac{\sum x^2}{49} - \frac{50 \times 40^2}{49}$ or

$\sum x^2 = (484 \times 49) + (50 \times 40^2) = 1,03,716$

Correct sample:

Sum of the observations, $\sum x = 2,000 - 150 + 50 = 1,900$

$\therefore \bar{x} = \frac{\sum x}{n} = \frac{1,900}{50} = 38$

$\therefore$ Correct variance,

$$\sigma^2 = \frac{\sum x^2}{n-1} - \frac{n\bar{x}^2}{n-1}$$

$$= \frac{83,716}{49} - \frac{50 \times 38^2}{49} = 235.02$$

$\therefore$ Correct standard deviation = $\sqrt{\sigma^2} = 15.33$ (approx.)

Illustration 10

Compute the mean deviation from mean and median from the following data:

Marks	10	20	30	40	50	60
No. of Student	4	7	15	8	7	2

Solution

Calculation of Mean deviation from median

Marks	Frequency	cf	|D|	f|D|
10	4	4	20	80
20	7	11	10	70
30	15	26	0	0
40	8	34	10	80
50	7	41	20	140
60	2	43	30	60
	N= 43			$\sum$ f|D| = 430

Median = Size of $\frac{N+1}{2}$ th item = $\frac{43+1}{2} = 22^{nd}$ item.

The size of 22^{nd} item is 30.

MD = $\frac{\sum |D|}{N} = \frac{430}{43} = 10$

Coefficient of MD = $\frac{MD}{Median} = \frac{10}{30} = 0.333$

Calculation of Mean Deviation from Mean

Marks	Frequency	fx	|D|	f|D|
10	4	40	23	92
20	7	140	13	91
30	15	450	3	45
40	8	320	7	56
50	7	350	17	119
60	2	120	27	54
	N = 43	$\sum$ fx = 1420		$\sum$ f|D| = 457

$$\bar{X} = \frac{\sum fx}{N} = \frac{1,420}{43} = 33$$

$$MD = \frac{\sum |D|}{N} = \frac{457}{43} = 10.63$$

$$\text{Coefficient of MD} = \frac{MD}{\text{Median}} = \frac{10.63}{33} = 0.322$$

SUMMARY

- Measures of dispersion give us an idea about the sensitivity of data, which the measures of Central Tendency does not reveal.
- Range is the simplest measure of dispersion and it is the difference between the highest and the lowest data point, but it is not a reliable measure as it is influenced by the largest and the smallest values in the data. On the other hand, deviation measures take into account every data.
- The Mean Absolute Deviation computes the absolute value of deviation of the data points from the mean. Apart from this, inter-quartile range, quartile deviation and deciles are other measures of dispersion.
- Among all the measures of dispersion, standard deviation is the most appropriate measure of dispersion as it overcomes the problem of the positive and negative deviations canceling each other, by squaring the deviations, then finding the mean of the squared deviation and then taking the positive square root of the mean of squared deviation.
- Variance is the average squared deviation from the arithmetic mean.
- The standard deviation is usually regarded as the most powerful measure of dispersion. It is free from those defects suffered by other measures. It lends itself to the analysis of variability in terms of a normal curve of error.
- Practically, all advanced statistical methods deal with variability and center around the standard deviation. Hence, unless the circumstances warrant the use of any other measure, the standard deviation can be used for measuring variability.

Chapter VIII

Skewness

After reading this chapter, you will be conversant with:

- Types of Distributions
- Meaning of Skewed Distribution
- Measures of Skewness
- Additional Illustrations

Introduction

The measures of central tendency and dispersion studied earlier do not reveal the entire story of a frequency distribution. The measures of central tendency describe the concentration of the observations about the center of the distribution. The measures of dispersion describe the scatteredness or spread of the observations about the measures of central tendency. The two comparable characteristics that help in understanding a distribution are Skewness and Kurtosis. These two distributions will have the same mean and standard deviation but differ widely from one another in its overall appearance. Though the two frequency distributions have same mean = 15, and the standard deviation = 6, yet they widely differ in shape and size when presented on the histogram, it is a skewed distribution. The present chapter deals with Skewness and its related measures.

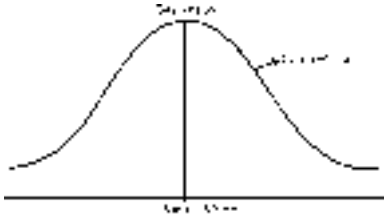
TYPES OF DISTRIBUTIONS

The following can be regarded as different types of skewed distributions:

Symmetrical Distribution

The value of mean, mode and median coincide in a symmetrical distribution. In such distribution, the frequency is equally spread on both the sides of the center point of the curve. In other words, the two halves coincide with each other and consequently both the tails, i.e. right and left of the curve will be equal in shape and length. The interval between mean and median is approximately one-third of the interval between the mean and mode. It will be clear from the following diagrams:

Figure 1: Symmetrical Distribution



Source: Gupta S.P. 'Statistical Methods' Pg no. 331.

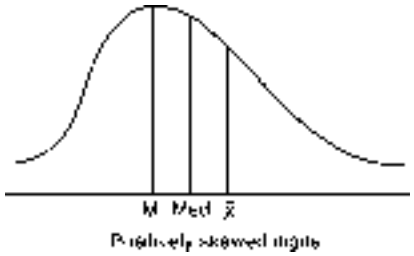
Asymmetrical Distribution

A distribution which is not symmetrical is called Skewed distribution. The frequency curve of such a distribution is not symmetric (bell shaped curve). The value of mean, mode and median will not coincide and the quartiles (Q_1 and Q_3) are not equidistant from the median. Similarly, the pairs of deciles and percentiles are also not equidistant from the median and the sum of positive and negative deviations from the median are not equal. The curve stretches more to one side than to the other side of the curve i.e. one side will have a longer tail than the other side. Such skewed distribution could be either positively skewed or negatively skewed. The skewness lies between +3 or (–3), but it is a rare phenomenon.

a. **Positively Skewed Distribution:** The frequency curve of the distribution, which has a longer tail towards the right side of the curve is said to be positively skewed distribution. In other words, greater range of the values are spread on the high-value end than the low-value end. In a positively skewed distribution, the value of mean is maximum, the value of mode is minimum

or least, and the median lies in between the two values i.e. mean and mode. It is clear from the following diagram:

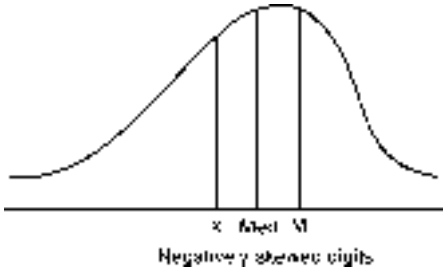
Figure 2: Positively Skewed Distribution



Source: Gupta S.P. 'Statistical Methods' Pg no. 331.

b. **Negatively Skewed Distribution:** The frequency curve of the distribution, which has a longer tail towards the left side of the curve is said to be Negatively skewed distribution. In such distribution, the greater variation is towards the lower value of the variables. In the negatively skewed distribution, the value of mean is least or minimum, the value of mode is maximum and the median lies in between the two. The shape of the negatively skewed distribution will be as follows:

Figure 3: Negatively Skewed Distribution



Source: Gupta S.P. 'Statistical Methods' Pg no. 331.

Tests for Skewness

The following tests are applied for ascertaining whether a distribution is skewed or not. A distribution is said to be skewed or asymmetrical if:

- The data plotted on the graph will not give a normal bell-shaped form.
- The value of mean, mode and median are not equal.
- Quartiles Q_1 and Q_3 are not equidistant from the median.
- The sum of positive deviations from median is not equal to the sum of the negative deviations from the median.
- Corresponding pairs of deciles and percentiles are not equidistant from the median.
- Frequencies are not distributed equally at the equal deviation point from the mode.

MEANING OF SKEWED DISTRIBUTION

Meaning

In literal sense, Skewness refers to 'lack of symmetry'. In other words, when the distribution is not symmetrical or is asymmetrical the distribution is called skewed distribution. It gives an idea about the shape of the curve drawn with the help of given frequency distribution. Concentration of observation towards higher or lower value, its nature and extent is known with the help of skewness. If the frequency curve of a unimodal symmetrical frequency distribution is folded at the center, the two halves coincide with each other. As compared to normal distribution, the measure of skewness indicates the difference in items which are

distributed in a particular distribution. For example, a positive skewness indicates that the frequencies of the distribution are spread over the right hand side (high-value end) of the curve than on the left hand side. Whereas in normal distribution, the spread is equal on the both the sides of the center and the mean, median and mode will be same value. Some of the definitions are given below in order to have a clear meaning of skewness. It is of much use in analyzing performance of business, measuring inflation etc.

Definitions

According to Croxton and Cowden "When a series is not symmetrical it is said to be asymmetrical or skewed."

According to Morris Hamburg "Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution."

According to Garrett "A distribution is said to be 'skewed' when the mean and the median fall at different points in the distribution, and the balance (or center of gravity) is shifted to one side or the other – to the left or right."

Objectives of Skewness

- It helps in finding out the nature and degree of concentration of distribution viz., whether it lies in higher values or lower values i.e., it is a positively skewed data or negatively skewed data.
- It gives clarity on the empirical relationship between mean, median and mode. This empirical relationship is based on skewness.
- Normal distribution is the basis for many statistical measures such as error of mean etc. These measures are based on the assumptions of a normal distribution. As such, it should be looked into that the distribution is a normal distribution. Skewness is the measure used for this purpose.

Difference between Dispersion and Skewness

The distinction between Dispersion and Skewness can be studied from the following points:

Dispersion	Skewness
It refers to the spread of individual variable values around a central value in a distribution.	It signifies the symmetry of distribution values on both sides of the central value.
It is regarded as one of the types of average because it signifies the deviations around a central value.	It is not a type of average, but, it is measured on the basis of various types of average viz., mean, median and mode.
It exhibits the degree of variability of the distribution.	It helps in finding out whether the concentration of distibution is in higher values or in lower values.
It indicates the accuracy of mean of the data i.e., whether the mean represents all the values in the data.	It helps in judging if the distribution is normal.
It deals with general variability of the distribution.	It indicates the symmetry of distribution on either side of the mode.
It exhibits the general shape of frequency distribution.	It refers to the dispersion on both sides of the mode in arranging the frequencies.

MEASURES OF SKEWNESS

The extent and direction of asymmetry or lack of symmetry in a series is known with the help of different measures of skewness. It allows the comparison between two or more series in this regard. The measure of skewness may be absolute or relative.

Absolute Measures of Skewness

The absolute measure of skewness measures the difference between the mean and mode in absolute terms. It is symbolically represented as follows:

Absolute S_K = Mean – Mode

Or (based on empirical relationship)

Absolute S_K = Mean – Median

Or (based on empirical relationship)

Absolute S_K = Median – Mode

The distribution is a positive skewed distribution, when the value of mean is greater than the value of mode. The distribution is negatively skewed, when the value of mode is greater than the value of mean. The difference between mean and mode is taken for measuring the skewness, because in a symmetrical distribution the value of mean, mode and median are equal. However, in a asymmetrical distribution, the mean moves away from the mode. Therefore, mean and mode are taken for measuring the skewness – the greater is the distance the more asymmetrical distribution is.

LIMITATIONS

The absolute measure of skewness is not of much practical utility for the following reasons:

- It involves the units of measurement, which cannot be used for comparing the two distributions expressed in different units of measurements.
- The absolute measure is not recommended even if the distributions have the same unit of measurement. The reason being that the distributions may have identical skewness but differs widely in the measure of central tendency and dispersion.

Relative Measures of Skewness

As we know that the absolute measure cannot be used for comparing the two distributions expressed in different units. So for comparing the two or more distributions for skewness, we use the relative measure of skewness. The relative measures of skewness are also known as Coefficient of skewness which are independent of measurement. In this measure, we divide the absolute measure of skewness by a suitable measure of dispersion so that the dispersion is eliminated. This measure has following three properties:

- It should be a pure number i.e. it is independent of unit of measurement and also of degree of variations in the series.
- When the distribution is symmetrical, the measure will have a zero value.
- It should be capable of interpreting the measured value.

The following are the coefficients of skewness which are popularly used:

- Karl Pearson’s Coefficient of Skewness.
- Bowley’s Coefficient of Skewness.
- Kelly’s Coefficient of Skewness.

They have been studied as under:

KARL PEARSON'S COEFFICIENT OF SKEWNESS

This measure of skewness is given by Karl Pearson, a British Biometrician and Statistician. This method is popularly known as Pearsonian coefficient of Skewness and denoted as Sk_P. It is the difference between the mean and mode divided by standard deviation. This is given by the following formula:

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

The value of coefficient of skewness lies between ±1. In a symmetrical distribution, the value of mean, mode and median coincide and so the coefficient of skewness is zero. The coefficient of skewness will have positive sign, when the distribution is positively skewed and will have a minus sign, when the distribution is negatively skewed.

When the mode is ill-defined, it is difficult to locate it. So in such case, for a moderately asymmetrical distribution we use the empirical relationship between the mean, median and mode i.e.,

Mode = 3 Median – 2 Mean

By subsitituting this value of mode in the above formula, we get

$$Sk_p = \frac{[\bar{X} - 3(\text{Med} - 2\bar{X})}{\sigma} = \frac{\bar{X} - 3\text{Med} + 2\bar{X}}{\sigma} = \frac{3(\bar{X} - \text{Med})}{\sigma}$$

Illustration 1

Calculate the Karl peason's coefficient of skewness

X	2.5	7.5	12.5	22.5	17.5	27.5	32.5	37.5
f	28	42	54	108	129	61	45	33

Solution

Calculation of Karl Pearson's Coefficient of Skewness

X	f	(X – 17.5)/5 = d	d^2	fd	fd^2
2.5	28	–3	9	–84	252
7.5	42	–2	4	–84	168
12.5	54	–1	1	–54	54
17.5	108	0	0	0	0
22.5	129	+1	1	+129	129
27.5	61	+2	4	+122	244
32.5	45	+3	9	+135	405
37.5	33	+4	16	+132	528
	N = 500			$\sum$ fd = 296	$\sum fd^2$ = 1780

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

$$\bar{X} = A + \frac{\sum fd}{N} \times c$$

A = 17.5, $\sum$ fd = 296; N = 500 and C = 5

$$= \quad 17.5 + \frac{296}{500} \times 5 = 20.46$$

Mode, by inspection is 22.5, as the maximum frequency is 129.

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N}\right]^2} \times c$$

$\sum fd = 296$, $N = 500$, $\sum fd^2 = 1780$, and $C = 5$

$$\sigma = \sqrt{\frac{1780}{500} - \left[\frac{296}{500}\right]^2} \times 5$$

$$= \sqrt{3.56 - .35} \times 5 = 8.96$$

Coefficient of Skewness

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{20.46 - 22.5}{8.96} = -0.228$$

Comment: The distribution is negatively skewed.

Illustration 2

Calculate Karl Pearson's coefficient of skewness from the data given below:

Weekly Wages (Rs.)	No. of Workers (f)
30-40	5
40-50	6
50-60	8
60-70	10
70-80	25
80-90	30
90-100	36
100-110	50
110-120	60
120-130	70

Solution

Calculation of Karl Peason's Coefficient of Skewness

Weekly Wages (Rs.)	No. of Workers (f)	Mid-value (X)	$d = \frac{X - 75}{10}$	fd	d^2	fd^2	c.f
30-40	5	35	–4	–20	16	80	5
40-50	6	45	–3	–18	9	54	11
50-60	8	55	–2	–16	4	32	19
60-70	10	65	–1	–10	1	10	29
70-80	25	75	0	0	0	0	54
80-90	30	85	+1	+30	1	30	84
90-100	36	95	+2	+72	4	144	120
100-110	50	105	+3	+150	9	450	170
110-120	60	115	+4	+240	16	960	230
120-130	70	125	+5	+350	25	1750	300
	N = 300			$\sum fd = 778$		$\sum fd^2 = 3510$	

In this illustration, as the highest frequency 70 is occurring at the end, the mode is ill-defined. So, the Karl Pearson's coefficient of skewness is calculated by using median.

$$Sk_p = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

$$\bar{X} = A + \frac{\sum fd}{N} \times c$$

A = 75, $\sum fd = 778$; N = 300 and C = 10

$$= 75 + \frac{778}{300} \times 10 = 100.93$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N}\right]^2} \times c$$

$\sum fd = 778$, N = 300, $\sum fd^2 = 3510$, and C = 10

$$= \sqrt{\frac{3510}{300} - \left[\frac{778}{300}\right]^2} \times 10$$

$$= \sqrt{11.7 - 6.724} \times 10 = 22.307$$

Median = N/2th item = 300/2 = 150 th item lies in the class of 100 – 110. So the median

$$\text{Median} = L + \frac{N/2 - Cf}{f} \times i$$

$$\text{Median} = 100 + \frac{150 - 120}{50} \times 10$$

$$= 100 + \frac{30}{50} \times 10$$

$$= 100 + 6 = 106$$

$$Sk_p = \frac{3(100.93 - 106)}{22.304}$$

$$= \frac{3 \times (-5.07)}{22.307} = -0.682$$

Comment: It is a negatively skewed distribution.

BOWLEY'S COEFFICIENT OF SKEWNESS

This measure was proposed by Professor A.L. Bowley, which is based on quartiles. It is also known as Quartiles Coefficient of Skewness. As said earlier, first and third quartiles are equidistant from median in a symmetrical distribution whereas in an asymmetrical distribution, the third quartile is the same distance over the median and the first quartile is found below the median.

Q_3 – Median = Median – Q_1 or

$Q_3 + Q_1 - 2\text{Median} = 0$

It is calculated by following formula:

$$Sk_b = \frac{(Q_3 - \text{Median}) - (\text{Median} - Q_1)}{(Q_3 - \text{Median}) + (\text{Median} - Q_1)}$$

$$\Rightarrow \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

It is not possible to compare the results obtained by these two measures (Pearson and Bowley) with one another because of computational basis that limits its value to +1 and –1. It is also possible that one may give positive skewness and the other may give a negative skewness.

However, Bowley's coefficient of skewness is useful,

- if presence of extreme observations in the data and mode is ill-defined.
- in case of open end classes or un-equal class intervals.

Illustration 3

Calculate Bowley's coefficient of skewness for the following frequency distribution.

Families	0	1	2	3	4	5	6
No. of Children's	7	10	16	25	18	11	8

Solution

Calculation of Bowley's Coefficient of Skewness

Families (x)	No. of Children (f)	cf
0	7	7
1	10	17
2	16	33
3	25	58
4	18	76
5	11	87
6	8	95

$$Sk_b = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Median = Size of $\frac{N+1}{2}$ th item = $\frac{95+1}{2} = 48\text{th item}$

The size of 48th item is 3. Therefore, Median = 3.

Q_1 = Size of $\frac{N+1}{4}$ th item = $\frac{95+1}{4} = 24\text{th item}$

The size of 24th item is 2. Hence, $Q_1 = 2$

Q_3 = Size of $\frac{3(N+1)}{4}$ th item = $\frac{3(95+1)}{4}$ = 72nd item

hence Q_3 is 4.

$$Sk_b = \frac{4+2-2(3)}{4-2} = 0$$

Comment: Since the coefficient of skewness is 0, it's been treated that there exists no asymmetry and is a symmetrical data.

Illustration 4

The following table gives the distribution of monthly incomes of 500 workers in a factory. Calculate Bowley's Coefficient of skewness:

Monthly Income (Rs.)	No. of Workers
Below Rs.100	10
100-150	25
150-200	145
200-250	220
250-300	70
300 and above	30

Solution

Calculation of Bowley's Coefficient of Skewness

Monthly income	No. of Workers (f)	cf
Below-100	10	10
100-150	25	35
150-200	145	180
200-250	220	400
250-300	70	470
300 and above	30	500

$$Sk_b = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Median = Size of $\frac{N}{2}$ th item = $\frac{500}{2} = 250$ th item, which lies in class 200-250.

Here, L = 200, N/2 = 250, cf =180, f = 220 and i = 50

$$\text{Med} = L + \frac{N/2 - cf}{f} \times i$$

$$= 200 + \frac{250 - 180}{220} \times 50$$

$$= 200 + \frac{70}{220} \times 50$$

$$= 200 + 15.909 = 215.91$$

Q_1= Size of $\frac{N}{4}$ th item = $\frac{500}{4} = 125$th item , which lies in 150-200 class

Here, L = 150, N/4 = 125, cf = 35, f = 145 and i = 50

$$Q_1 = L + \frac{N/4 - cf}{f} \times i$$

$$= 150 + \frac{125 - 35}{145} \times 50$$

$$= 150 + \frac{90}{145} \times 50 = 150 + 31.03 = 181.03$$

Q_3 = Size of $\frac{3N}{4}$ th item = $\frac{3(500)}{4} = 375$th item which lies in 200-250 class

Here, L = 200, 3N/4 = 375, cf =180, f = 220 and i = 50

$$Q_3 = L + \frac{3N/4 - cf}{f} \times i$$

$$= 200 + \frac{375 - 180}{220} \times 50$$

$$= 200 + \frac{195}{220} \times 50$$

$$= 200 + 44.32 = 244.32$$

$$Sk_b = \frac{244.32 + 181.03 - 2(215.91)}{244.32 - 181.03}$$

$$= \frac{425.35 - 431.82}{63.29} = \frac{-6.47}{63.29} = -0.1022$$

Illustration 5

From the following data, you are required to calculate coefficient of skewness based on quartiles and median:

Marks	No. of Students
0-10	12
10-20	16
20-30	26
30-40	38
40-50	22
50-60	15
60-70	7
70-80	4

Solution

The Bowley's coefficient of Skewness given by the following formula is based on quartiles and median:

$$Sk_B = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Calculation of Bowley's Coefficient of Skewness

Marks	f	c.f.
0-10	12	12
10-20	16	28
20-30	26	54
30-40	38	92
40-50	22	114
50-60	15	129
60-70	7	136
70-80	4	140

Q_1 = Size of $\frac{N}{4}$ th item = $\frac{140}{4} = 35$ th item. It lies in the class 20-30.

$$Q_1 = L + \frac{N/4 - cf}{f} \times i$$

L = 20, N/4 = 35, cf = 28, f = 26, i = 10

$$Q_1 = 20 + \frac{35 - 28}{26} \times 10 = 20 + 2.69 = 22.69$$

Q_3 = Size of $\frac{3N}{4}$ th item = $\frac{3\times140}{4}$ =105th item. It lies in the class 40-50.

$$Q_3 = L + \frac{3N/4 - cf}{f} \times i$$

L = 40, 3N/4 = 105, cf = 92, f = 22, i = 10

$$Q_3 = 40 + \frac{105 - 92}{22} \times 10 = 40 + 5.91 = 45.91$$

$$\text{Median} = \text{Size of } \frac{N}{2} \text{ th item} = \frac{140}{2} = 70\text{th item. It lies in the class 30-40}$$

$$\text{Median} = L + \frac{N/2 - cf}{f} \times i$$

L = 30,N/2 = 70, cf = 54,f = 38, i = 10

$$\text{Median} = 30 + \frac{70 - 54}{38} \times 10 = 30 + 4.21 = 34.21$$

Coefficient of

$$Sk_B = \frac{45.91 + 22.69 - 2(34.21)}{45.91 - 22.69} = \frac{68.6 - 68.42}{23.22} = 0.0008 \approx 0.$$

Comment: The data is a positively skewed data.

Kelly's Coefficient of Skewness

The drawback of Bowley's measure is that it ignores the two extremes of the data. This drawback can be eliminated partially by taking two deciles or percentiles equidistant from the median. So, it is extension of Bowley coefficient based on quartiles. This refinement was suggested by Kelly which base its formula for measuring skewness on 10th and 90th percentiles or 1st and 9th deciles. The formula is given below:

$$Sk_k = \frac{P_{90} + P_{10} - 2\text{Median}}{P_{90} - P_{10}} \text{ also}$$

$$Sk_k = \frac{D_9 + D_1 - 2\text{Median}}{D_9 - D_1}$$

This method is seldom used in practice and is of theoretical importance.

ADDITIONAL ILLUSTRATIONS

Illustration 1

From the following calculate Karl Pearson's coefficient of skewness

Age	No. of Persons
20-25	5
25-30	7
30-35	8
35-40	18
40-45	15
45-50	12
50-55	7
55-60	5

Solution

Calculation of Karl Pearson's Coefficient of Skewness

Age	No. of persons	m	(m – 37.5)/5 = d	d^2	fd	fd^2
20-25	5	22.5	–3	9	–15	45
25-30	7	27.5	–2	4	–14	28
30-35	8	32.5	–1	1	–8	8
35-40	18	37.5	0	0	0	0
40-45	15	42.5	+1	1	+15	15
45-50	12	47.5	+2	4	+24	48
50-55	7	52.5	+3	9	+21	63
55-60	5	57.5	+4	16	+20	80
	N = 77				$\sum fd = 43$	$\sum fd^2 = 287$

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

$$\bar{X} = A + \frac{\sum fd}{N} \times c$$

A = 37.5, $\sum fd = 43$; N = 77 and C = 5

$$= 37.5 + \frac{43}{77} \times 5 = 40.29$$

Since it is difficult to know the modal class by inspection, we have to prepare analysis and grouping table.

Grouping Table

Age	I	II	III	IV	V	VI
20-25	5					
		12				
25-30	7					
			15			
30-35	8					
		26				
35-40	**18**					
			33	20	23	**41**
40-45	15					
		27				
45-50	12					
			19			
50-55	7					
		12		**45**	**34**	24
55-60	5					

Analysis Table

Age	I	II	III	IV	V	VI	Total
35-40	1		1	1		1	4
40-45	1	1	1	1	1		5
45-50	1		1	1			3

From grouping and analysis table, it is clear that the modal class is 40-45

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \qquad \Delta_1 = (18-15); \Delta_2 = (15-12)$$

$$= 40 + \frac{3}{3+3} \times 5$$

$$= 40 + \frac{3}{6} \times 5 = 40 + 2.5 = 42.5$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N}\right]^2} \times c$$

$\sum fd = 43$, $N = 77$, $\sum fd^2 = 287$, and $C = 5$

$$= \sqrt{\frac{287}{77} - \left[\frac{43}{77}\right]^2} \times 5$$

$$= \sqrt{3.727 - 0.312} \times 5 = 9.24$$

$$Sk_p = \frac{40.29 - 42.5}{9.24} = \frac{-2.21}{9.24} = -0.24$$

Illustration 2

From the information given below, calculate Karl Pearson's coefficient of Skewness.

	X	Y
Mean	150	120
Median	135	15
Standard Deviation	135	15

Do both the distributions have the same degree of variation and skewness (Sk_p)?

Solution

a. **Degree of Variation**

$$CV = \frac{\sigma}{\bar{X}} \times 100$$

$$\text{For (X)} = CV = \frac{15}{150} \times 100 = 10\%$$

$$\text{For (Y)} = CV = \frac{15}{135} \times 100 = 11.11\%$$

The Y distribution is more variable than distribution X.

b. **Degree of Skewness**

$$Sk_p = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

$$\text{For (X)} = \frac{3(150-135)}{15} = 3$$

$$\text{For (Y)} = \frac{3(135-120)}{15} = 3$$

The degree of skewness is the same for both the distributions.

Illustration 3

For a distribution, Karl pearson's coefficient of skewness is 0.5 and mode is 26, median is 32. Calculate the coefficient of variation.

Solution

$$Sk_p = \frac{Mean - Mode}{\sigma}$$

We are given Pearson's coefficient = 0.5, mode = 26, and median = 32

Mode = 3Median – 2 Mean

26 = 3(32) – 2 Mean

26 = 96 – 2 Mean

2Mean = 96 – 26

Mean = 70/2 = 35.

$$0.5 = \frac{35 - 26}{\sigma}$$

$$0.5\sigma = 9$$

$$\sigma = \frac{9}{0.5} = 18$$

$$CV = \frac{18}{35} \times 100 = 51\% \text{ (approx….)}$$

Illustration 4

In the frequency distribution, the coefficient of skewness based on quartiles is 0.5. The sum of upper and lower quartiles is 100 and the median is 38. Calculate the value of upper and lower quartiles.

Solution

$$Sk_b = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$$

given Median = 38, Sk_b= 0.50, $Q_3 + Q_1$= 100

$$0.50 = \frac{100 - 2(38)}{Q_3 - Q_1}$$

$$Q_3 - Q_1 = \frac{100 - 76}{0.50} = 48$$

$Q_3 + Q_1 = 100$

$Q_3 - Q_1 = 48$

$2Q_3 = 52$

$Q_3 = 52/2 = 26$ and

$Q_1 = 100 - 26 = 78$

Illustration 5

Calculate Karl Pearson and Bowleys coefficient of skewness from the following data:

Profits	10-20	20-30	30-40	40-50	50-60
No. of firms	27	30	45	33	15

Solution

Calculation of Karl Pearson's Coefficient of Skewness

Profits	No. of Firms	m	(m – 35)/10 = d	d^2	fd	fd^2	cf
10-20	27	15	–2	4	–54	108	27
20-30	30	25	–1	1	–30	30	57
30-40	45	35	0	0	0	0	102
40-50	33	45	+1	1	+33	33	135
50-60	15	55	+2	4	+30	60	150
	N = 150				$\sum fd = -21$	$\sum fd^2 = 231$	

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

$$\bar{X} = A + \frac{\sum fd}{N} \times c$$

A = 35, $\sum fd = -21$; N = 150 and C = 10

$$= 35 + \frac{-21}{150} \times 10 = 33.6$$

By inspection, the Modal class is 30-40

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i ; \qquad \Delta_1 = (45 - 30); \Delta_2 = (45 - 33)$$

$$= 30 + \frac{15}{15 + 12} \times 10$$

$$= 30 + \frac{15}{27} \times 10$$

$$= 30 + 5.5 = 35.5$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left[\frac{\sum fd}{N}\right]^2} \times c$$

$\sum fd = -21$, N = 150, $\sum fd^2 = 231$, and C = 10

$$\sigma = \sqrt{\frac{231}{150} - \left[\frac{-21}{150}\right]^2} \times 10$$

$$= \sqrt{1.54 - 0.0196} \times 10 = 12.33$$

$$Sk_p = \frac{33.6 - 35.5}{12.23} = \frac{-1.9}{12.23} = -0.155$$

Bowleys coefficient of skewness:

$$Sk_b = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$$

Median = Size of $\frac{N}{2}$ th item = $\frac{150}{2} = 75$ th item, which lies in class 30-40.

$$Med = L + \frac{N/2 - cf}{f} \times i$$

$$= 30 + \frac{75 - 57}{45} \times 10$$

$$= 30 + \frac{18}{45} \times 10$$

$$= 30 + 4 = 34$$

Q_1= Size of $\frac{N}{4}$ th item = $\frac{150}{4} = 37.5$th item

$$Q_1 = L + \frac{N/4 - cf}{f} \times i$$

$$= 20 + \frac{37.5 - 27}{30} \times 10$$

$$= 20 + \frac{10.5}{30} \times 10$$

$$= 20 + 3.5 = 23.5$$

$$Q_3 = \text{Size of } \frac{3N}{4} \text{ th item} = \frac{3(150)}{4} = 112.5\text{th item}$$

$$Q_3 = L + \frac{3N/4 - cf}{f} \times i$$

$$= 40 + \frac{112.5 - 102}{33} \times 10$$

$$= 40 + \frac{10.5}{33} \times 10$$

$$= 40 + 3.182 = 43.182$$

$$Sk_b = \frac{43.182 + 23.5 - 2(34)}{43.182 - 23.5}$$

$$= \frac{66.682 - 68}{19.682} = \frac{-1.318}{19.682} = -0.067$$

It is a negatively skewed distribution.

Illustration 6

Calculate Bowley's Coefficient of Skewness from the following data:

Annual Sales	Less than	10	20	30	40	50	60	70
No. of Firms		12	30	60	75	84	89	90

Solution

Calculation of Bowley's Coefficient of Skewness

Annual Sales	No. of Firms (f)	cf
0-10	12	12
10-20	18	30
20-30	30	60
30-40	15	75
40-50	9	84
50-60	5	89
60-70	1	90

$$Sk_b = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Median = Size of $\frac{N}{2}$ th item = $\frac{90}{2} = 45$ th item, which lies in class 20-30

$$\text{Med} = L + \frac{N/2 - cf}{f} \times i$$

$$= 20 + \frac{45 - 30}{30} \times 10$$

$$= 20 + \frac{15}{30} \times 10$$

$$= 20 + 5 = 25$$

Q_1= Size of $\frac{N}{4}$ th item = $\frac{90}{4} = 22.5$th item which lies in class 10-20

$$Q_1 = L + \frac{N/4 - cf}{f} \times i$$

$$= 10 + \frac{22.5 - 12}{18} \times 10$$

$$= 10 + \frac{10.5}{18} \times 10$$

$$= 10 + 5.83 = 15.83$$

$$Q_3 = \text{Size of } \frac{3N}{4} \text{ th item} = \frac{3(90)}{4} = 67.5\text{th item}$$

$$Q_3 = L + \frac{3N/4 - cf}{f} \times i$$

$$= 30 + \frac{67.5 - 60}{15} \times 10$$

$$= 30 + \frac{7.5}{15} \times 10 \quad = \ 30 + 5 = 35$$

$$Sk_b = \frac{35 + 15.83 - 2(25)}{35 - 15.83}$$

$$= \frac{50.83 - 50}{19.17}$$

$$= \frac{+0.83}{19.17} \ = \ +0.0433$$

SUMMARY

- Skewness refers to lack of symmetry which occurs in case of an unsymmetrical distribution for which the mean, median and mode are not equal.
- For a positively skewed distribution: AM > median > mode.
- For a negatively skewed distribution: AM < median < mode.
- The extend and direction of asymmetry in a series is known with the help of measure of skewness. The measure of skewness may be absolute or relative measure. The absolute measure measures the difference between mean and mode. The relative measure is known as coefficient of skewness, which is independent of measurement.

Chapter IX

Correlation

After reading this chapter, you will be conversant with:

- Meaning and Definition of Correlation
- Types of Correlation
- Significance of Correlation
- Methods of Studying Correlation
- Karl Pearson's Coefficient of Correlation
- Coefficient of Correlation and Probable Error
- Rank Correlation Coefficient
- Coefficient of Determination
- Concurrent Deviation Method
- Additional Illustrations

Introduction

There are many instances where managers take decisions based on future events. For this, they rely on observations of two or more variables, which appear to be related to one another. However, it is not appropriate to make major business decisions depending on such vague evidence without investigations concerning the form and strength of any relationship that exists.

The Coefficient of Correlation which gives a measure of the relevance of the model is discussed hereunder:

The board of directors of ABC Company is facing a problem of estimating what the annual sales might be in a shop to be opened in Bagpur where ABC has not operated before. They need this information in order to plan the size of the shop, the amount of stock to be put on the showcase, the number of employees to be hired. An answer to this problem may be found statistically by establishing the general relationship in the cities in which the company is already operating, say between the size of a city's employed labor force or working population and sales in its ABC shops. From the size of Bagpur's employed labor force, an estimation of the annual sales in that new shop can be estimated, and the board of directors should be able to base its decision on this estimate.

However, the board will also want to know how good this estimate is, since it is based only on a general relationship between sales and employment. Further, the board may want to compare the relationship existing between sales and labor force on one hand with the relationship existing between sales and number of shops on the other hand. Such problems can be solved through correlation analysis.

MEANING AND DEFINITION OF CORRELATION

Correlation does not deal with one series, but rather with the association or relationship between two series and does not measure variation in one series, but rather compare variation in two or more series. The existence of correlation between variables does not necessarily mean that one is the cause of the movement in the other. It should be noted that the correlation analysis merely helps in determining the degree of association between two variables, but it does not tell anything about the cause and effect relationship.

Correlation analysis is based on the relationship between two or more variables. The degree of relationship between the variables under consideration is measured through the correlation analysis. The measure of correlation called correlation coefficient, summarizes in one figure the direction and degree of correlation. Correlation analysis is a technique that measures the closeness of the relationship between the variables such as prices of commodities and the amount demanded, age of father and son, etc.

Definition

According to Croxton and Cowden "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as Correlation".

According to Ya-Lun-Chou "Correlation analysis attempts to determine the degree of relationship between variables."

According to A.M. Tuttle "Correlation is an analysis of the covariation between two or more variables."

From the above definitions, it is clear that Correlation is a Statistical device, which studies the relationship between the two variables. The Correlation analysis consists of various methods and techniques that can be used for studying and measuring the extent of relationship between the two variables. If the change in one variable leads to a corresponding change in other variable, then the two variables are said to be correlated. The direction of change is indicated by + or – signs; the former refers to the sympathetic movement in the same direction and the latter, in the opposite direction; an absence of correlation is indicated by zero. Thus, the coefficient of correlation ranges between –1 and 1.

TYPES OF CORRELATION

The three important types of correlation are:

1. Positive and Negative Correlation.
2. Linear and Non-linear Correlation.
3. Simple, Partial and Multiple Correlation.

Positive and Negative Correlation

Positive and Negative Correlation depends on the direction of change of the variables. If the value of two variables changes/deviates in the same direction i.e., if the increase in the value of one variable leads on an average to a corresponding increase in the value of another variable or if the decrease in the value of one variable leads to a decrease in the value of other variable on an average, the correlation is said to be *Positive* or *Direct*. Example of such correlations are price and supply for the commodity, rainfall and yield of crop, the income and expenditure of a family etc.

On the other hand, if the variables deviates/vary in opposite direction i.e. if the increase in the value of one variable leads to a decrease in the value of other variable on an average, such a correlation is known as *Negative* or *Inverse* Correlation. Examples of negative correlations are price and demand for a product, sale of umbrella and the winter season etc.

The difference between the positive and negative correlation can be more clearly understood with the help of the following illustration:

Positive Correlation

Price (X)	20	30	40	50	60
Supply (Y)	30	40	50	60	70

Negative Correlation

Price (X)	20	30	40	50	60
Demand (Y)	60	50	40	30	20

Linear and Non-linear Correlation

The distinction between linear and non-linear correlation is based on the constancy of ratio of change between the two variables. If the amount of change/unit of change in the value of one variable leads to constant amount of change in the other variable over the entire range of the values, then the correlation is said to be Linear correlation. Let take an example to illustrate it.

X	1	2	3	4	5
Y	7	14	21	28	35

Thus, from the above example, it is clear that a unit change in the value of X will lead to a constant change in the corresponding value of Y. The ratio of change between the two variables is same. If such a data is plotted on the graph as points on xy-plane, then all the points fall on a straight line.

If the amount of change/unit of change in the value of one variable does not change the value of other variable at constant rate but at fluctuating rate, such a correlation is known as Non-Linear or Curvilinear Correlation. For example, if the amount rainfall is doubled, the yield of crop may not necessarily get doubled. If such data is plotted on the xy-plane, the we do not get straight line curve. Practically, a non-linear relationship exists between the variables that are commonly found in economics and social sciences. Since the techniques used for analysis and measurement of non-linear correlation are complicated and tedious when compared to linear correlation, we generally assume that the relationship between the variable is linear. The difference between the linear and non-linear correlation can be clearly understood with the help of diagrams referred to in later pages.

Simple, Partial and Multiple Correlation

This distinction is based on the number of variables to be studied. The correlation is said to be simple correlation when only two variables are studied. The simple correlation is also known as bivariate distribution because the each unit of series assumes two values only. It is Multiple or Partial correlation when we study three or more variables. In multiply correlation, we study three or more variables simultaneously, whereas in partial correlation we consider more than two variables. It is also known as multivariate distribution.

SIGNIFICANCE OF CORRELATION

The study of correlation is of immense use in practical life because there exists some kind of relationship between the variables. The degree of relationship that exists between the variables, is studied with the help of correlation analysis. If the two variables are closely related, then we can estimate the value of one variable from the given value of another variable with the help of regression analysis. Correlation analysis also contributes to the understanding of economic behavior, aids in locating the critically important variables on which other variables depend, may reveal to the economist the connection by which disturbances spread and suggest to him the paths through, which stabilizing forces may become effective.

In business, correlation analysis enables the executive to estimate costs, sales, prices and other variables on the basis of some other series with which these costs, sales, or prices may be functionally related.

Some of the guesswork can be removed from decisions when the relationship between a variable to be estimated and the one or more variables on which it depends are close and reasonably invariant.

However, it should be noted that coefficient of correlation is one of the most widely used and also most widely abused statistical measures. It is abused in the sense that one sometimes overlooks the fact that correlation measures nothing but the strength of linear relationship and that it does not necessarily imply a cause-effect relationship.

METHODS OF STUDYING CORRELATION

The various methods for ascertaining the linear relationship or correlations between the two variables are as follows:

1. Scatter Diagram method.
2. Graphic Method.
3. Karl Pearson's Coefficient of Correlation.
4. Rank Method.
5. Concurrent Method.

The first two methods are the methods that are based on diagram and graph whereas the other methods are based on mathematical methods.

Scatter Diagram Method

The first step in correlation analysis is to visualize the relationship. For each unit of observation in correlation analysis, there is a pair of numerical values. One is considered the independent variable; the other variable is depended on it and is called the dependent variable. One of the easiest ways of studying the correlation between the two variables is with the help of a scatter diagram. It is a diagrammatic representation of a bivariate distribution. Scalter diagram is a simple tool for ascertaining whether the two variables are correlated or not by preparing a dot chart. The dot chart is known as Scatter diagram. The given data is plotted on a graph paper in the form of dots (.) i.e., we put a dot for each pair of X and Y on the x-axis and y-axis in the xy-plane.

A scatter diagram give us two types of information. Visually, we can look for patterns that indicate whether the variables are related. Then, if the variables are related, we can see what kind of line, or estimating equation, describes this relationship.

The scatter diagram gives an indication of the nature of the potential relationship between the variables.

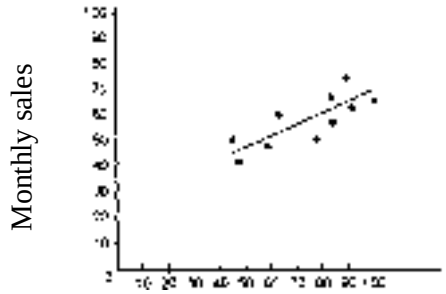
Illustration 1

A sample of 10 employees of the Universal Computer Corporation was examined to relate the employees' score on an aptitude test taken at the beginning of their employment and their monthly sales volume. The Universal Computer Corporation wishes to estimate the nature of the relationship between these two variables.

Aptitude Test Score	Monthly Sales (in Thousands of Rupees)
X	Y
50	30
50	35
60	40
60	50
70	55
70	60
70	45
80	55
80	50
90	65

Solution

To determine the nature of the relationship for example, we initially draw a graph to observe the data points.



Aptitude test score

On the vertical axis, we plot the dependent variable monthly sales. On the horizontal axis, we plot the independent variable aptitude test score. This visual display is called a scatter diagram.

In the figure given above, we see that larger monthly sales are associated with larger test scores. If we wish, we can draw a straight line through the points plotted in the figure. This hypothetical line enables us to further describe the relationship.

A line that slopes upward to the right indicates that a direct, or a positive relation is present between the two variables. In the figure given above, we see that this upward-sloping line appears to approximate the relationship being studied.

The figures below show additional relations that may exist between two variables. In figure (a), the nature of the relationship is linear. In this case, the line slopes downward. Thus, smaller values of Y are associated with larger values of X. This relation is called an inverse (linear) relation.

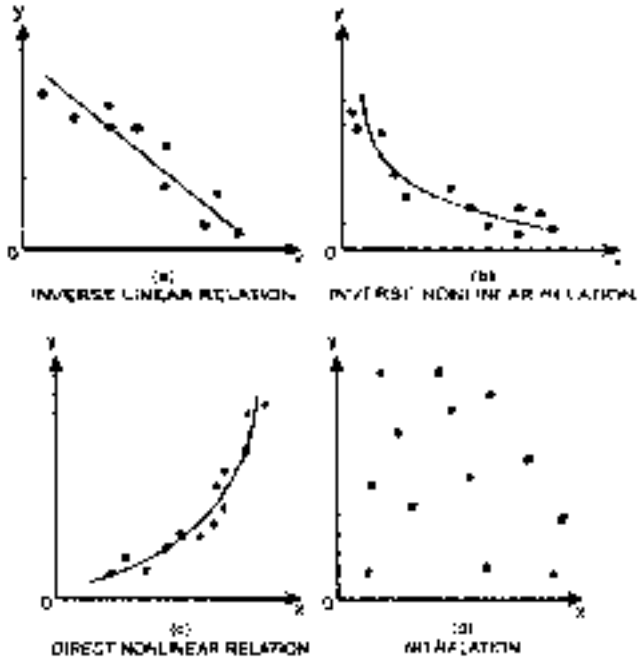


Figure (b) represents a relationship that is not linear. The nature of the relationship is better represented by a curve than by a straight line – that is, it is a curvilinear relation. The relationship is inverse since smaller values of Y are associated with larger values of X.

Figure (c) is another curvilinear relation. In this case, however, larger values of Y are associated with larger values of X. Hence, the relation is direct and curvilinear.

In figure (d), there is no relation between X and Y. We can draw neither a straight line nor a curve that adequately describes the data. The two variables are not associated.

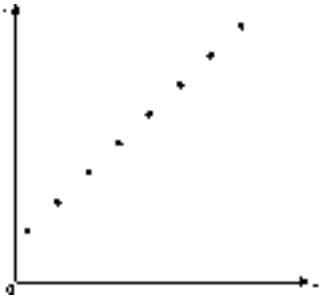
A scatter diagram gives an idea whether the variables are related to each other or not. The interpretation that can be drawn from the scatter diagram is given below:

- If the points plotted on the graph are widely scattered, then there is lesser or poor correlation between the two variables. If the points are closely related to each other, then a good correlation is expected between the variables see figure (a).
- If the scatter diagram reveals a trend then the variables are said to be correlated and the variables are uncorrelated if the trend is not revealed.
- The correlation is said to be positive (i.e., the value of two variables moves in same direction) if the trend is rising upward from the lower left hand corner to the upper right hand corner [see figure b]. The correlation is said to be negative (i.e., the value of two variables moves in opposite direction) if the trend is downward from the upper left hand corner to the lower right hand corner [see figure c].

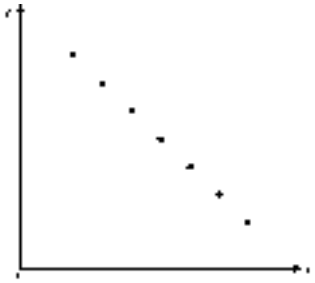
- The correlation is said to be perfectly positive (r = +1), if all points lie on the straight line falling from the bottom left hand corner to the upper right-hand corner. The correlation is said to be perfectly negative (r = –1) if all the points lie on a straight line rising from the upper left corner to the lower right hand.

The following figures illustrate the different types of Correlation:

Perfect Positive Correlation

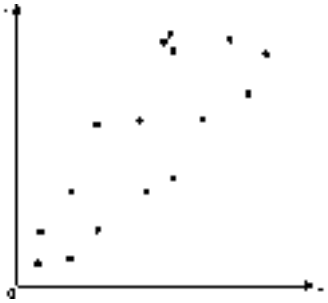


Perfect Negative Correlation

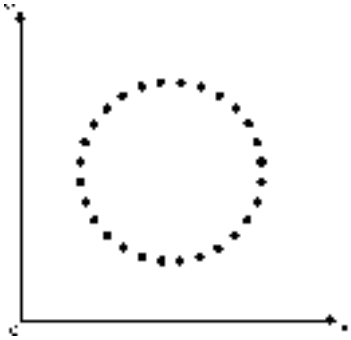


High Degree of Negative Correlation

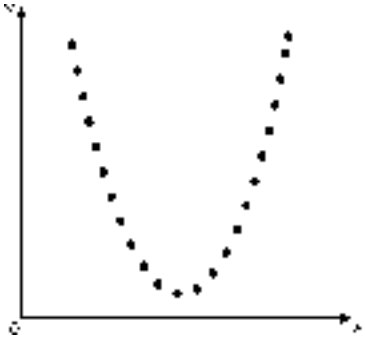
Low Degree of Positive Correlation



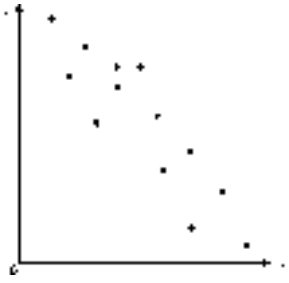
No Correlation



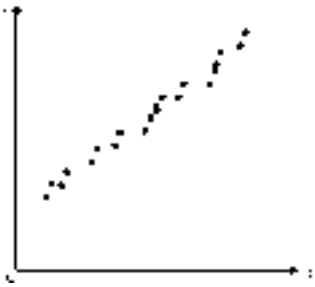
No Correlation



Low Degree of Negative Correlation



High Degree of Positive Correlation



Source: Adapted from Arora P. N. and Arora S., Statistics (C.A., Foundation Course).

Advantages and Limitations of Scatter Diagram

Advantages:

i. It is a simple non-mathematical method for studying the correlation between the two variables.

ii. This method provides the information relating to the nature of relationship, but not the extent of relationship that exists between the variables.

iii. Scatter diagrams are not affected by the extreme observations whereas most of the statistical methods of ascertaining the correlation are affected by the extreme observation or by the size of extreme items.

iv. The scatter diagram provides the line of best fit by free hand method i.e., drawing a line through the plotted points to locate the best possible line.

Limitation: The only drawback of this method is that it does not give the exact degree of correlation between the two variables.

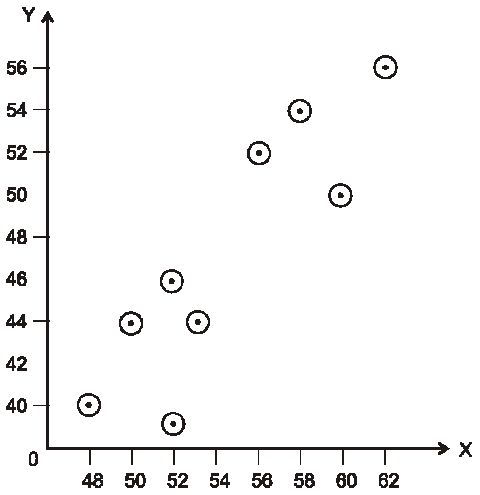
Illustration 2

Following are the values of two variables – height and weight of students in a class:

Height (in inches) X	52	62	58	48	60	55	52	50
Weight (in kgs) Y	40	55	53	40	44	50	51	44

Draw a scatter diagram and interpret the correlation between them.

Solution



As the points are close to each other, we may expect a high degree of correlation between the two variables X and Y. secondly, as the points show an upward trend from the left bottom towards right top, the correlation is positive between height and weight of students.

Graphic Method

In graphic method, we plot the individual values of two variables on the graph paper and obtain two curves i.e. one for X variable and another for Y variable. To know whether the variables are closely related or not, we need to examine the direction and closeness of the curves drawn. The correlation is said to be positive when both the curves move in same direction i.e., either upward or downward. If

both the curves move in the opposite/inverse direction, then the correlation is said to be negative. This method is followed when the data is given for a period of time. Let us take an illustration to have a clear understanding:

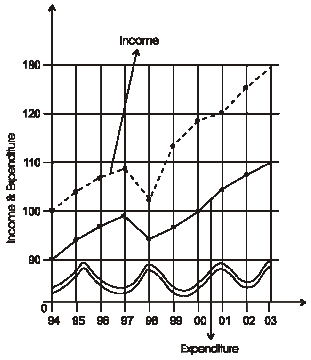
Illustration 3

Following are the two variables relating to 50 workers of a factory. Ascertain whether there is any correlation between the income and expenditure of workers.

Year	Annual Income (Rs.)	Annual Expenditure (Rs.)
1994	2100	2000
1995	2320	2200
1996	2500	2350
1997	2700	2500
1998	3200	3000
1999	3300	3100
2000	3500	3200
2001	3650	3300
2002	3800	3650
2003	4000	3500

Solution

Figure



From the above graph, it is clear that the two variables – Income and Expenditure of factory workers are closely related to each other.

KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl Pearson's coefficient of correlation is one of the several mathematical methods available for measuring the correlation i.e., the magnitude of linear relationship between the two variables. Karl Pearson, a British Biometrician and Statistican suggested and is popularly known as Pearson's coefficient of correlation which is denoted by (r). It is the most widely used method in practice. The symbol 'r' is the numerical measure of linear relationship between the two variables. It is also defined as the covariance between the two variables (Cov (xy)) to the product of standard deviation of two variables. The value of 'r' always lies between ± 1. There is perfectly positive correlation between two variables, when the value of $r = +1$ and when the value of $r = -1$, then the correlation between the two variables is perfectly negative. No relationship exists between the two variables when the value of $r = 0$. The correlation coefficient is exists calculated by the following formula:

$$r = \frac{\sum xy}{n\sigma_x\sigma_y} \text{ or } \frac{Cov.(x,y)}{\sigma_x\sigma_y} \text{ or } \frac{\frac{1}{n}\sum xy}{\sigma_x\sigma_y}$$

Where,

$x = (X - \bar{X})$; $y = (Y - \bar{Y})$.

σ_x = Standard Deviation of series X.

σ_y = Standard Deviation of series Y.

n = Number of pairs of observation.

r = Coefficient of correlation.

When the deviations are taken from actual mean, the above formula is used. The above formula can be simplified as follows to avoid the tedious calculations:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Where,

$x = (X - \bar{X})$ and $y = (Y - \bar{Y})$

The following steps should be followed for calculating the correlation coefficient:

1. Calculate the mean of X and Y variables.
2. Obtain the deviations of X series from the mean of X, denote it by x.
3. Square the deviations so obtained and take the total i.e., $\sum x^2$
4. Adopt similar procedure for Y series.
5. Multiply the deviation of X and Y series and sum it up i.e. $\sum xy$
6. Substitute the values in the formula for obtaining the 'r'.

Illustration 4

The following data relates to the sales and net profit of 10 chemical companies for the year ended March '99. Let us calculate the coefficient of correlation from the following data and interpret its value.

(Rs. in crore)

Name of the Company	Sales	Net Profits
Aegis Chemical Industries	37.72	0.19
Alembic Chemical Works	168.13	1.60
Alkyl Amines Chemicals Ltd.,	17.69	1.30
Citurgia Biochemicals Ltd.,	45.21	2.62
Colour Chem Ltd.,	159.52	5.32
Dharamsi Morarji Chemicals	151.79	6.52
Diamines and Chemicals Ltd.,	23.27	1.05
Godavari Fertilizers and Chemicals	442.84	10.31
Gujarat Alkalies and Chemicals	191.54	16.68
Herdillia Chemicals Ltd.,	149.16	6.78

Solution

Let sales be denoted by X and the net profits by Y.

Calculation of Co-efficient of Correlation

	X	$(X-\bar{X})$ $= x$	$(X-\bar{X})^2$ $= x^2$	Y	$(Y-\bar{Y})$ $= y$	$(Y-\bar{Y})^2$ $= y^2$	xy
Aegis Ind.	37.72	–100.97	10,194.94	0.19	–5.05	25.05	509.09
Alembic	168.13	29.44	866.71	1.60	–3.64	13.25	–107.16
Alkyl Ltd.,	17.69	–121.00	14,641.00	1.30	–3.94	15.52	476.74
Citurgia Ltd.,	45.21	–93.48	8,738.51	2.62	–2.62	6.86	244.92
Colour Chem.	159.52	20.83	433.89	5.32	0.08	0.006	1.67
DM Chem	151.79	13.10	171.61	6.52	1.28	1.64	16.77
Diamines	23.27	–115.42	13,321.78	1.05	–4.19	17.56	483.61
Godavari	442.84	304.15	92,507.22	10.31	5.07	25.71	1,542.04
G Alkalies	191.54	52.85	2,793.12	16.68	11.44	130.87	604.60
Herdillia	149.16	10.47	109.62	6.78	1.54	2.37	16.12
Total	1386.87		1,43,778.40	52.37		239.29	3,789.21

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

$$\bar{X} = \frac{\sum X}{N} = \frac{1386.87}{10} = 138.687;$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{52.37}{10} = 5.24$$

$$= \frac{3789.21}{\sqrt{143778.4 \times 239.29}} = 0.65$$

The positive correlation coefficient (r = 0.65) evaluated above indicates a reasonably high degree of association between the variables X and Y or more precisely between sales and net profit. It indicates that higher the sales, higher would be the net profit. It shows that the sales and net profits move in the same direction.

When Deviations are Taken from Assumed Mean: When the actual mean is in fractions, i.e., in illustration no.3. then the computation of coefficient of correlation by actual mean becomes tedious and time-consuming. So in such case, we can use the assumed mean method i.e., taking the deviations from assumed mean rather than the actual mean for calculating the correlation coefficient. The formula is given below:

$$r = \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sqrt{\sum dx^2 - \frac{(\sum d_x)^2}{N} \times \sum d_y{}^2 - \frac{(\sum d_y)^2}{N}}}$$

Where,

d_x $= (X - A)$; $d_y = (Y - A)$

$\sum d_x$ = Sum of the deviations of X taken from assumed mean (X – A)

$\sum d_y$ = Sum of the deviations of Y taken from assumed mean (Y – A).

$\sum d_x^2$ = Sum of the square of deviations of X taken from assumed mean.

$\sum d_y^2$ = Sum of the square of deviations of Y taken from assumed mean.

$\sum d_x d_y$ = Sum of the product of deviations of X and Y taken from assumed mean.

The following steps are followed when the deviations are taken from assumed mean:

1. Obtain the total of deviations taken from assumed mean of X series and denote it by $\sum d_x$.
2. Obtain the total of deviations taken from assumed mean of Y series and denote it by $\sum d_y$.
3. Square the deviations taken from assumed mean of X series, obtain the total and denote it by $\sum d_x^2$.
4. Square the deviation taken from assumed mean of Y series, obtain the total and denote it by $\sum d_y^2$.
5. Multiply the deviations obtained from assumed mean of X series with the deviations of Y series, obtain the total and denote it by $\sum d_x d_y$.
6. Substitute the values in the formula and calculate the value of 'r'.

Illustration 5

Calculate the coefficient of correlation between the price and demand for a product from the following data. Assume 59 and 102 as the mean value for price and demand respectively.

Price	68	79	89	50	49	69	58	51
Demand	115	127	146	102	97	126	113	98

Solution

Calculation of Coefficient of correlation

X	$(X - 59) = d_x$	d_x^2	Y	$(Y-102) = d_y$	d_y^2	$d_x\ d_y$
68	+9	81	115	+13	169	+117
79	+20	400	127	+25	625	+500
89	+30	900	146	+44	1936	+1320
50	–9	81	102	0	0	0
49	–10	100	97	–5	25	+50
69	+10	100	126	+24	576	+240
58	–1	1	113	+11	121	–11
51	–8	64	98	–4	16	+32
$\sum x = 513$	$\sum d_x = 41$	$\sum d_x^2 = 1727$	$\sum y = 924$	$\sum d_y = 108$	$\sum d_y^2 = 3468$	$\sum d_x d_y = 2248$

$$r = \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sqrt{\sum dx^2 - \frac{(\sum d_x)^2}{N} \times \sum d_y^2 - \frac{(\sum d_y)^2}{N}}}$$

$$= \frac{(2248)-(41)(108)/8}{\sqrt{1727-\frac{(41)^2}{8}\times 3468-\frac{(108)^2}{8}}}$$

$$= \frac{2248-553.5}{\sqrt{1727-\frac{1681}{8}\times 3468-\frac{11664}{8}}}$$

$$= \frac{1694.5}{\sqrt{1727-210.125\times 3468-1458}}$$

$$= \frac{1694.5}{\sqrt{1516.875\times 2010}} = \frac{1694.5}{\sqrt{3048918.75}} = \frac{1694.5}{1746} = +0.97$$

Assumption of Karl Pearson's Coefficient of Correlation: The assumptions on which Karl Pearson's coefficient of correlations based are:

- It assumes that the relationship between the variables is linear.
- Large number of independent causes affects the variable under study before forming a normal distribution.
- There should be cause and effect relationship between the two variables. If there is no such relationship or if the variables are independent, then there is no correlation.

Merits and Limitations of Pearsonian Coefficient of Correlation

Karl Pearson method is most popular for measuring the degree of relationship between the variables. It describes not only the magnitude of correlation, but also the direction. However, it also suffers from certain limitations:

- The first assumption of linear assumption is not correct.
- Secondly, the coefficient of correlation should be interpreted with great care and its value is unduly affected by the extremes. It is a time-consuming method.

Interpretations of Coefficient of Correlation

Correlation Analysis is performed to measure the degree of association between two variables. The measure is called coefficient of correlation. The coefficient of correlation is also said to be a measure of covariance between two series. Utmost care must be taken while interpreting the value of correlation coefficient as the reliability of estimates depends on closeness of relationship. The investigator should thoroughly understand the data so as to avoid the misinterpretions. The following rule will help in interpreting the value of 'r':

Degree of Correlation	Positive	Negative
Absence of Correlation	Zero	Zero
Perfect Correlation	+1	– 1
High degree	+0.75 to +1	– 0.75 to –1
Moderate degree	+ 0.25 to + 0.75	– 0.25 to – 0.75
Low degree	0 to + 0.25	0 to – 0.25

The correlation coefficient describes not only the magnitude of correlation, but also its direction.

Properties of Coefficient of Correlation

The properties of correlation of coefficient are given below:

1. The coefficient of correlation always lies between –1 and +1. Symbolically, it is represented as $-1 \leq r \leq +1$.
2. The coefficient of correlation is independent of change origin and scale of the two variables X and Y.

3. The geometric mean of two regression coefficients is equal to coefficient of correlation.
4. The degree of relationship between two variables is symmetric.

COEFFICIENT OF CORRELATION AND PROBABLE ERROR

The value of coefficient of correlation can be interpreted with the help of probable error. It helps in determining the reliability of estimates of the value of coefficient as it depends on the conditions of random sampling. The probable error of coefficient of correlation is denoted as $(P.E_r)$ calculated as follows:

$$P.E_r = 0.6745\frac{1-r^2}{\sqrt{N}}$$

Where,

r = coefficient of correlation.

N = Number of pairs of observations.

- There is no correlation between the variables, if the value of r is less than the probable error.
- The value of r is significant, if its value is 6 times more than the probable error.
- The upper and lower limit of coefficient of correlation of the population can be determined by adding and deducting the probable error value from the coefficient of correlation $r \pm P.E.$

The Probable error utility depends on the following three conditions:

1. The data must approximate a normal frequency curve.
2. The statistical measure for which probable error is computed must have been calculated from a sample.
3. The sample must have been selected in an unbiased manner and the individual items must be independent.

Illustration 6

If r = 0.5 and N = 36 find, the probable error of the coefficient of correlation and determine the limit for population r.

Solution

$$P.E_r = 0.6745\frac{1-r^2}{\sqrt{N}}$$

$$P\,E_r = 0.6745\frac{1-(0.5)^2}{\sqrt{36}} = \frac{0.6745\times 0.75}{6} = 0.084$$

Limit for population correlation $= 0.5 \pm 0.084 = 0.584$ to 0.416.

RANK CORRELATION COEFFICIENT

When the population under study is not normally distributed or when the shape of distribution is not known, we cannot adopt the Karl Pearson's method for computing the coefficient of correlation. So in such case, we need a measure of correlation, which does not make any assumption about the population parameter. In such measure, we can rank the observations according to size and make the calculation on the basis of rank rather than the original observations. This measure gives us the coefficient of rank correlation. Charles Edward Spearman, a British Psychologist in 1904, developed the rank correlation as a measure for finding out the covariability or lack of covariability between the two variables. It is especially useful in those cases, when the characteristics whose possible correlation is being investigated, cannot be measured but individuals can only be ranked on the basis of the characteristics to be measured. We then have two sets of ranks available for working out the correlation coefficient. Sometimes, the data on one variable may be in the form of ranks while the data on the other variable is in the form of measurements which can be converted into ranks.

Thus, when both the underlying variables are ordinal or when the data are available in the ordinal form irrespective of the type of variable, we use the rank correlation coefficient to find out the extent of relationship between the two variables.

This rank correlation coefficient is also known as Spearman's rank correlation coefficient. Rank correlation is a non-parametric technique for measuring the strength of relationship between paired observations of two variables when the data is in a ranked form. It is denoted as R.

Rank Correlation measures the degree of agreement between the two sets of ranks.

Rank Correlation Coefficient R is

$$R = 1 - \frac{6\sum_{i=1}^{n} D^2}{n^3 - n}$$

where,

n = Number of individuals ranked.

D = Difference in the ranks of the i^{th} individual.

i = 1, 2, …….. n.

Features of Spearman's Correlation Coefficient

1. $\sum D = 0$ i.e., the sum of the difference of rank between the two variables shall be zero.
2. Spearman's correlation coefficient does not make any assumption about the population from which the sample is drawn, i.e., non-parametric.
3. The interpretation of spearman's correlation coefficient is similar to Pearsonian correlation coefficient.

The rank correlation deals with two types of problem:

- When ranks are given.
- When the ranks are not given.
- When equal ranks are given.

WHEN RANKS ARE GIVEN

Illustration 7

The ranks of 8 students are given according to their marks in English and History.

Student No.	1	2	3	4	5	6	7	8
English	2	8	1	7	6	4	5	3
History	7	3	2	6	5	8	4	1

We want to know whether or not students who are good at English are also good at History. For this, the calculations are:

n	Difference between Ranks (D)	D^2
1	–5	25
2	5	25
3	–1	01
4	1	01
5	1	01
6	–4	16
7	1	01
8	2	04
Total	0	74

Here $\Sigma D^2 = 74$ and n = 8

$$R = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 74}{512 - 8} = 0.119$$

The correlation between the ranks is seem to be low.

WHEN RANKS ARE NOT GIVEN

When the ranks are not assigned to the actual data, we need to assign the ranks to it. The ranks are assigned either by giving the highest value as 1 or the lowest value as 1, but the same method should be followed in both the variables.

Illustration 8

Calculate spearman's coefficient of correlation between the marks assigned by A and B in a certain competitive exam as shown below:

Marks by A	42	43	32	50	35	31	27	28	15	17
Marks by B	55	58	33	28	67	38	25	20	15	40

Solution

We first assign the rank – 1 for the lowest and then calculate the rank correlation coefficient.

Marks by A	R_x	Marks by B	R_y	$(R_x - R_y) = D^2$
42	8	55	8	0
43	9	58	9	0
32	6	33	5	1
50	10	28	4	36
35	7	67	10	9
31	5	38	6	1
27	3	25	3	0
28	4	20	2	4
15	1	15	1	0
17	2	40	7	25
				$\sum D^2 = 76$

$$R = 1 - \frac{6\Sigma D^2}{n^3 - n}$$

$$R = 1 - \frac{6 \times 76}{10^3 - 10}$$

R = 1 – 0.461 = 0.539

WHEN EQUAL RANKS ARE GIVEN

In certain cases, the ranks of individuals are equal. In such a case, we assign average rank to each individual. Suppose, two individuals are ranked equal to the third place, then they are given rank = $\frac{3+4}{2} = 3.5$ and if three are ranked equal to the third place, then the ranks assigned to them are $\frac{3+4+5}{3} = 4$.

The above formula for calculating the rank correlation coefficient is adjusted when equal ranks are assigned. The formula is

$$R = 1 - \frac{6\{\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + ...\}}{N^3 - N}$$

$\frac{1}{12}(m^3 - m)$ is added to the value of $\sum D^2$ as an adjustment when equal ranks are allotted to the variables. 'm' stands for the items whose ranks are common. $\frac{1}{12}(m^3 - m)$ is added to the value of $\sum D^2$ if there are more than one common rank.

Illustration 9

Calculate Spearman's rank Correlation Coefficient for the variables X and Y.

X	40	45	55	40	45	50	40	55	60	65
Y	100	100	105	115	130	105	120	110	105	150

Solution

First assign the ranks –1 for the lowest, 2 for the next highest, 3 for the next highest and so on.

Calculation of Spearman's Correlation Coefficient

X	R_x	Y	R_y	$(R_x - R_y) = D^2$
40	2	100	1.5	0.25
45	4.5	100	1.5	9.00
55	7.5	105	4	12.25
40	2	115	7	25.00
45	4.5	130	9	20.25
50	6	105	4	4.00
40	2	120	8	36.00
55	7.5	110	6	2.25
60	9	105	4	25.00
65	10	150	10	0
				$\sum D^2 = 134$

Note: In Series X 40 is repeated three times (m = 3), 45 is repeated twice (m = 2), and 55 is repeated twice (m = 2) and in Y series, 100 is repeated twice, (m = 2), and 105 is repeated thrice (m = 3).

$$R = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)\right\}}{N^3 - N}$$

$$R = 1 - \frac{6\left\{134 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right\}}{10^3 - 10}$$

$$R = 1 - \frac{6\{134 + 2 + 0.5 + 0.5 + 0.5 + 2\}}{990}$$

$$= 1 - \frac{6(139.5)}{990} = 1 - \frac{837}{990}$$

$$= 1 - 0.845 = 0.155$$

Merits and Limitations of Rank Correlation Coefficient

Merits:

- When compared to the Karl Pearson's method, this method is easy and simple to understand and apply. Both the methods will given the answer if the items are not repeated.
- It can be used advantageously, when the data is of qualitative nature.
- This is the only method which does not use actual data but assigns rank to it.

Limitations:

- We cannot compute the correlation for a grouped frequency distribution with this method.
- The method is time-consuming and tedious when the number of items exceeds 30.

However, this method can be used in the following cases:

- When the actual data is in the form of ranks
- When the number of items is small, say not more than 25 or 30, then rank correlation can be applied as an approximation to time-consuming Pearsonian's Correlation Coefficient.

COEFFICIENT OF DETERMINATION

The coefficient of determination is given by r^2 i.e., the square of the correlation coefficient. It explains to what extent the variation of a dependent variable is expressed by the independent variable. A high value of r^2 shows a good linear relationship between the two variables. If r = 1 and r^2 = 1, it indicates a perfect relationship between the two variables.

CONCURRENT DEVIATION METHOD

Concurrent deviation is the simplest method. In this method, we have to find out the direction of change of variables X and Y. It is denoted as r_c and calculated by the following formula:

$$r_c = \pm\sqrt{\pm\left[\frac{2C-n}{n}\right]}$$

Where, r_c = Coefficient of correlation by Concurrent method.

C = Sum of positive signs obtained after multiplying D_x and D_y.

n = Number of pairs of observation = N – 1.

The following steps will be made for calculating the correlation coefficient by Concurrent method:

- Ascertain the direction of change of X and Y variables i.e., compare the second value with the first value. If the second value is increasing put a '+' sign and put '–' sign if the value is decreasing and 0 when the two values are same/constant. Repeat this procedure for other values and denote the X series as D_x and follow the same procedure for Y series and denote it by D_y.
- Multiply D_x with D_y to determine the value of C and obtain its total, i.e., the total number of Positive signs.
- Apply the formula for calculating the correlation coefficient by this method.

Illustration 10

Calculate the correlation coefficient for the variables X and Y by Concurrent Deviation method.

X	40	45	55	40	45	50	40	55	60	65
Y	100	100	105	115	130	105	120	110	105	150

Solution

Calculation of Correlation coefficient by Concurrent Deviation method

X	D_x	Y	D_y	D_xD_y
40		100		
45	+	100	0	0
55	+	105	+	+
40	–	115	+	–
45	+	130	+	+
50	+	105	–	–
40	–	120	+	–
55	+	110	–	–
60	+	105	–	–
65	+	150	+	+
				C = 3

$$r_c = \pm\sqrt{\pm\left[\frac{2C-n}{n}\right]} = \pm\sqrt{\pm\left[\frac{2\times3-9}{9}\right]} = \sqrt{\frac{-3}{9}} = 0.577$$

Merits and Limitations of Concurrent Deviation method

Merits:

1. It is very simple method when compared to other methods used for computing the correlation.
2. Concurrent deviation method is used when the number of items is vary large.
3. It gives quick idea about the degree of relationship between the variables without using any complicated method.

Limitations:

1. This method roughly indicates the presence or absence of correlation.
2. All the values whether big or small are given equal weights when they vary in same direction.

Correlation of Grouped Data

When the number of observations is large, the data is often classified into two-way frequency distribution which is called a correlation table.

The class intervals for Y are listed in the captions or column headings, and those for X are lasted if the stubs at the left of the table, the order can also be reversed. The frequencies for each cell of the table are determined by either tallying or card sorting just as in the case of frequency distribution of a single variable.

The formula for calculating the coefficient of correlation is:

$$r = \frac{N\sum fd_Xd_Y - \sum fd_X fd_Y}{\sqrt{N\sum fd^2{}_X - (\sum fd_X)^2}\sqrt{N\sum fd^2{}_Y - (\sum fd_Y)^2}}$$

Steps:

- Take the step deviations of the variable X and denote these deviations by d_X.
- Take the step deviations of the variable Y and denote these deviations by d_Y.
- Multiply d_Xd_Y and respective frequency of each cell and write the figure obtained in right hand upper corner of each cell.

- Add together all the cornered values as calculated in step (iii) and obtain the total $\sum fd_X d_Y$.
- Multiply the frequencies of the variable X by the deviations of X and obtain the total $\sum fd_X$.
- Take the squares of the deviations of variable Y and multiply them by the respective frequencies and obtain $\sum fd^2{}_X$.
- Multiply the frequencies of the variable Y by the deviations of Y and obtain the total $\sum fd_Y$.
- Take the squares of the deviations of the variable Y and multiply them by the respective frequencies and obtain $\sum fd^2{}_Y$.
- Substitute the values of $\sum fd_X d_Y$, $\sum fd_X$, $\sum fd^2{}_X$, $\sum fd_Y$ and $\sum fd^2{}_Y$ in the above formula and obtain the value of r.

Calculation of Coefficient of Correlation

X		Age in years							
Y	d_X d_Y	16	17	18	19				
		–1	0	+1	+2	f	fd_Y	$fd^2{}_Y$	$fd_X d_Y$
100-200	–1	5 |_	0 |_	–3 |_	–4 |_	15	–15	15	–2
200-300	0	5 |_0	5 |_0	3 |_0	2 |_0	18	0	0	0
300-400	+1	4 |_–3	6 |_0	5 |_9	3 |_12	25	25	25	18
400-500	+2	3 |_–4	7 |_0	9 |_14	6 |_44	25	50	100	54
Total		2 14	5 23	7 24	11 22	N = 83	$\sum fd_Y$ = 60	$\sum fd^2{}_Y$ 140 =	$\sum fd_X d_Y$ = 70
fd_X		–14	0	24	44	$\sum fd_X$ =54			
$fd^2{}_X$		14	0	24	88	$\sum fd^2{}_X$ =126			
$fd_X d_Y$		–2	0	20	52	$\sum fd_X d_Y$ =70			

$$r = \frac{N\sum fd_X d_Y - \sum fd_X \sum fd_Y}{\sqrt{N\sum fd^2{}_X - (\sum fd_X)^2}\sqrt{N\sum fd^2{}_Y - (\sum fd_Y)^2}}$$

$$r = \frac{83\times 70 - 54\times 60}{\sqrt{83\times 126 - (54)^2}\ \sqrt{83\times 140 - (60)^2}}$$

$$= \frac{5810 - 3240}{\sqrt{10458 - 2916}\ \sqrt{11,620 - 3600}}$$

$$= \frac{2570}{\sqrt{7542}\ \sqrt{8020}}$$

$$= \frac{2570}{86.844\times 89.554} = \frac{2570}{7777.266} = 0.3304$$

ADDITIONAL ILLUSTRATIONS

Illustration 1

Consider the following sample of paired observations:

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Find the covariance between X and Y and use it to find the coefficient of correlation.

Solution

Calculation of Coefficient of Correlation

X	$(X-\bar{X})$ $= x$	$(X-\bar{X})^2$ $= x^2$	Y	$(Y-\bar{Y})$ $= y$	$(Y-\bar{Y})^2$ $= y^2$	xy
1	–6	36	1	–4	16	24
3	–4	16	2	–3	9	12
4	–3	9	4	–1	1	3
6	–1	1	4	–1	1	1
8	1	1	5	0	0	0
9	2	4	7	2	4	4
11	4	16	8	3	9	12
14	7	49	9	4	16	28
56		132	40		56	84

$$\bar{X} = \frac{\sum X}{N} = \frac{56}{8} = 7;\ \bar{Y} = \frac{\sum Y}{N} = \frac{40}{8} = 5$$

$$Cov_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1} = \frac{84}{7} = 12$$

$$\sigma_x = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}} = \sqrt{\frac{132}{8-1}} = \sqrt{18.857} = 4.34$$

$$\sigma_y = \sqrt{\frac{\sum(Y-\bar{Y})^2}{n-1}} = \sqrt{\frac{56}{8-1}} = \sqrt{8} = 2.83$$

$$r = \frac{Cov_{xy}}{\sigma_X \sigma_y} = \frac{12}{(4.34)(2.83)} = 0.977$$

Illustration 2

10 competitors in a beauty contest are ranked by 3 Judges A, B and C in the following order:

A	6	4	9	8	1	2	3	10	5	7
B	1	6	5	10	3	2	4	9	7	8
C	3	5	8	4	7	10	2	1	6	9

Compute the rank correlation coefficient for detemining the common taste of judges.

Solution

Computation of Spearman's Rank Coefficient of Correlation

R_1	R_2	R_3	$(R_1 - R_2)^2$ D^2	$(R_2 - R_3)^2$ D^2	$(R_1 - R_3)^2$ D^2
6	1	3	25	4	9
4	6	5	4	1	1
9	5	8	16	9	1
8	10	4	4	36	16
1	3	7	4	16	36
2	2	10	0	64	64
3	4	2	1	4	1
10	9	1	1	64	81
5	7	6	4	1	1
7	8	9	1	1	4
N = 10			60	200	214

Rank correlation between the judgment given by Judges A and B.

$$R = 1 - \frac{6\Sigma D_i^2}{n^3 - n}$$

$$R = 1 - \frac{6 \times 60}{10^3 - 10} = 1 - \frac{360}{990} = 0.636$$

Rank correlation between the judgment given by Judges B and C.

$$R = 1 - \frac{6\Sigma D_i^2}{n^3 - n}$$

$$R = 1 - \frac{6 \times 200}{10^3 - 10} = 1 - \frac{1200}{990} = 1 - 1.212 = -0.212$$

Rank correlation between the judgment given by Judges C and A.

$$R = 1 - \frac{6\Sigma D_i^2}{n^3 - n}$$

$$R = 1 - \frac{6 \times 214}{10^3 - 10} = 1 - \frac{1284}{990} = 1 - 1.297 = -0.297$$

Illustration 3

Calculate Spearman's coefficient of correlation from the following data between the price of tea and demand for tea:

Price of tea	65	78	85	60	50	70	71	40
Demand for tea	110	124	140	105	100	130	132	90

Solution

Calculation of Spearman's Correlation Coefficient

Price of Tea	R_1	Demand for Tea	R_2	$(R_1 - R_2)^2 = D^2$
65	4	110	4	0
78	7	124	5	4
85	8	140	8	0
60	3	105	3	0
50	2	100	2	0
70	5	130	6	1
71	6	132	7	1
40	1	90	1	0
				6

$$R = 1 - \frac{6\Sigma D_i^2}{n^3 - n}$$

$$R = 1 - \frac{6 \times 6}{8^3 - 8} = 1 - \frac{36}{504} = 1 - 0.071 = +0.929$$

Illustration 4

From the following data, calculate Karl Pearson's coefficient of correlation between the age and the players and also the probable error.

Age	20	22	25	30	38	40	42	45	47	51
Players	1	0	2	5	2	4	6	5	7	8

Solution

Calculation of Karl Pearson's Coefficient of Correlation

X	$(X - \bar{X}) = x$	$(X - \bar{X})^2 = x^2$	Y	$(Y - \bar{Y}) = y$	$(Y - \bar{Y})^2 = y^2$	xy
20	–16	256	1	–3	9	+48
22	–14	196	0	–4	16	+56
25	–11	121	2	–2	4	+22
30	–6	36	5	+1	1	–6
38	+2	4	2	–2	4	–4
40	+4	16	4	0	0	0
42	+6	36	6	+2	4	+12
45	+9	81	5	+1	1	+9
47	+11	121	7	+3	9	+33
51	+15	225	8	+4	16	+60
360	0	1092	40	0	64	230

$$\bar{X} = \frac{\sum X}{N} = \frac{360}{10} = 36; \ \bar{Y} = \frac{\sum Y}{N} = \frac{40}{10} = 4.$$

$$r = \frac{\sum XY}{\sqrt{\sum x^2 \times \sum y^2}}$$

$$= \frac{230}{\sqrt{1092 \times 64}} = \frac{230}{\sqrt{69888}} = \frac{230}{264.363} = +0.87$$

$$PE_r = 0.6745 \frac{1 - r^2}{\sqrt{N}}$$

$$PE_r = 0.6745 \frac{1 - (0.87)^2}{\sqrt{10}} = \frac{0.6745 \times 0.2431}{3.1623} = 0.05$$

Illustration 5

Calculate Karl Pearson's coefficient of correlation from the following data:

X	33	34	36	30	34	32	35	32	28	30	32	47
Y	19	21	9	8	9	17	17	19	31	20	16	10

Solution

Calculation of Karl Pearson's Coefficient of Correlation:

X	$(X-34)=d_x$	d_x^2	Y	$(Y-16)=d_y$	d_y^2	$d_x\ d_y$
33	–1	1	19	+3	9	–3
34	0	0	21	+5	25	0
36	+2	4	9	–7	49	–14
30	–4	16	8	–8	64	+32
34	0	0	9	–7	49	0
32	–2	4	17	+1	1	–2
35	+1	1	17	+1	1	+1
32	–2	4	19	+3	9	–6
28	–6	36	31	+15	225	–90
30	–4	16	20	+4	16	–16
32	–2	4	16	0	0	0
47	+13	169	10	–6	36	–78
	$\sum d_x = -5$	$\sum d_x^2 = 255$		$\sum d_y = 4$	$\sum d_y^2 = 484$	$\sum d_x d_y = -176$

$$r = \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sqrt{\sum dx^2 - \frac{(\sum d_x)^2}{N} \times \sum d_y^2 - \frac{(\sum d_y)^2}{N}}}$$

$$= \frac{(-176) - (5)(4)/12}{\sqrt{255 - \frac{(5)^2}{12} \times 484 - \frac{(4)^2}{12}}}$$

$$= \frac{-176 + 1.7}{\sqrt{255 - \frac{25}{12} \times 484 - \frac{16}{12}}}$$

$$= \frac{-174.333}{\sqrt{255 - 2.08 \times 484 - 1.33}}$$

$$= \frac{-174.333}{\sqrt{252.92 \times 482.67}} = \frac{-174.333}{\sqrt{122076.89}} = \frac{-174.333}{349.39} = -0.498$$

Illustration 6

A management consultant has collected the following data on the sales and advertising expenses of eight firms in the apparell industry:

Sales (Rs. in crore)	52	52	57	62	67	58	56	60
Advertising expenses (Rs. in crore)	11	13	16	17	18	14	15	16

Approximately what percentage of the variation in sales is explained by the variation in advertising expenses?

Solution

Let the following notations be used:

X	Advertising expenses (Rs. in crore)
Y	Sales (Rs. in crore)

Percentage of variation in Y that is explained by variation in X = Coefficient of determination = r^2

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} \quad \text{or} \quad \frac{\Sigma xy}{\Sigma x^2 \times \Sigma y^2}$$

Y	52	52	57	62	67	58	56	60	
X	11	13	16	17	18	14	15	16	
$Y-\bar{Y}=y$	–6	–6	–1	4	9	0	–2	2	
$X-\bar{X}=x$	–4	–2	1	2	3	–1	0	1	
$(Y-\bar{Y})^2=y^2$	36	36	1	16	81	0	4	4	178
$(X-\bar{X})^2=x^2$	16	4	1	4	9	1	0	1	36
$(X-\bar{X})(Y-\bar{Y})=xy$	24	12	–1	8	27	0	0	2	72

$$\bar{Y} = \frac{\sum X}{n} = \frac{464}{8} = 58 \qquad \bar{X} = \frac{\sum Y}{n} = \frac{120}{8} = 15$$

$$\therefore r = \frac{72}{\sqrt{36\times178}} = 0.8994 \quad \therefore r^2 = 0.8089 \approx 0.81$$

$\therefore$ The variation in advertising expenses explain 81% of the variation in sales.

Illustration 7

Classic Cosmetics Ltd., manufactures and sells cosmetic items. The following data has been collected from the company's finance department. The finance manager wants to know the degree of relationship between the sales and operating expenses of the company.

(Rs. in lakh)

Sales	92	98	80	107	112	128
Operating expenses	63	68	56	72	76	83

What is the standard error of estimate, which may occur if a regression line is plotted using the observations? Coefficient of correlation, r = 0.9955 and Total Sum of Squares = 457.33

Solution

$$\text{Coefficient of determination} = \frac{\text{Total sum of squares} - \text{Error sum of squares}}{\text{Total sum of squares}}$$

$$(0.9955)^2 = \frac{457.33 - ESS}{457.33}$$

$$ESS = 4.107$$

$$\text{Standard error estimate, } S_e = \sqrt{\frac{ESS}{n-2}} = \sqrt{\frac{4.107}{6-2}} = 1.013$$

Illustration 8

A sample of paired observations of two random variables, X and Y, is given below:

X	3	6	2	4	5	4	7	1
Y	25	32	16	28	30	13	30	10

What is the covariance between X and Y?

Solution

$$\text{Cov}(X,Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1} \text{ or } \frac{\Sigma xy}{n-1}$$

X	3	6	2	4	5	4	7	1	$\Sigma X = 32$
Y	25	32	16	28	30	13	30	10	$\Sigma Y = 184$
$X-\bar{X}=X$	–1	2	–2	0	1	0	3	–3	
$Y-\bar{Y}=Y$	2	9	–7	5	7	–10	7	–13	
xy	–2	18	14	0	7	0	21	39	$\Sigma xy = 97$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{32}{8} = 4$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{184}{8} = 23$$

$$\therefore \text{Cov}(X,Y) = \frac{97}{8-1} = 13.86 \text{ (approx).}$$

Illustration 9

A financial analyst has collected the following data on sales and net profits of five companies manufacturing the electronic components:

Sales (Rs. in crore)	38	18	45	24	55
Profit (Rs. in crore)	1.50	1.50	3.00	2.00	4.00

Calculate the coefficient of determination between sales and net profits.

Solution

Let the following notations be used:

X: Sales (Rs. in crore)

Y: Net profits (Rs. in crore)

$$\text{Coefficient of correlation, } r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

X	38.00	18.00	45.00	24.00	55.00	ΣX	=	180
Y	1.50	1.50	3.00	2.00	4.00	ΣY	=	12
$X-\bar{X}=X$	2.00	–18.00	9.00	–12.00	19.00			
$Y-\bar{Y}=Y$	–0.90	–0.90	0.60	–0.40	1.60			
xy	–1.80	16.20	5.40	4.80	30.40	Σxy	=	55
x^2	4.00	324.00	81.00	144.00	361.00	Σx^2	=	914
y^{27}	0.81	0.81	0.36	0.16	2.56	Σy^2	=	4.70

$$R = \frac{55}{\sqrt{914 \times 4.70}} \qquad \bar{x} = \frac{180}{5} = 36 \qquad \bar{y} = \frac{12}{5} = 2.4$$

$\therefore r = 0.839$

$\Rightarrow r^2 = 0.704$ (approx.)

Illustration 10

The following data pertains to the operations of Metro Finance Ltd., in the recent years:

Year	2002	2001	2000	1999	1998	1997
Amount of advances (Rs. in crore)	7.0	5.5	5.0	4.5	4.0	4.0
Profit after tax (Rs. in lakh)	54	48	42	27	19	14

If a simple regression equation is developed on the basis of the given data, with the amount of advances as the independent variable, then what will be the standard error of estimate?

Solution

Let the following notations be used:

X	Amount of advances (Rs. in crore)
Y	Profit after tax (Rs. in lakh)

Coefficient of correlation is given by:

$$r = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 (Y - \bar{Y})^2}}$$

X	7.0	5.5	5.0	4.5	4.0	4.0	$\sum X = 30$
Y	54	48	42	27	19	14	$\sum Y = 204$
$(X - \bar{X})$	2	0.5	0	–0.5	–1	–1	
$(Y - \bar{Y})$	20	14	8	–7	–15	–20	
$(X - \bar{X})^2$	4	0.25	0	0.25	1	1	$\sum (X - \bar{X})^2 = 6.5$
$(Y - \bar{Y})^2$	400	196	64	49	225	400	$\sum (Y - \bar{Y})^2 = 1334$
$(X - \bar{X})(Y - \bar{Y})$	40	7	0	3.5	15	20	$\sum (X - \bar{X})(Y - \bar{Y}) = 85.5$

$$\bar{X} = \frac{\sum X}{n} = \frac{30}{6} = 5.0$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{204}{6} = 34$$

$$\therefore r = \frac{85.5}{\sqrt{6.5 \times 1334}} = 0.9182$$

Proportion of variations in profit after tax that is explained by variations in the amount of advances = $r^2 = (0.9182)^2 = 0.8431$

$\therefore$ Percentage of variations in profit after tax that is explained by variations in the amount of advances = 0.8431 x 100 = 84.31%.

Proportion of variations in profit after tax that is explained by variations in the amount of advances

$$= r^2 = 1 - \frac{\Sigma (Y - \hat{Y})^2}{\Sigma (Y - \bar{Y})^2}$$

From the above, we have the following:

$r^2 = 0.8431$

$\Sigma(Y - \overline{Y})^2 = 1334$

$$\therefore 0.8431 = 1 - \frac{\Sigma(Y - \hat{Y})^2}{1334}$$

$$\text{or } \frac{\Sigma(Y - \hat{Y})^2}{1334} = 1 - 0.8431 = 0.1569$$

$\therefore \Sigma(Y - \hat{Y})^2 = 209.305$

Standard error of estimate

$$= \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}} = \sqrt{\frac{209.305}{6 - 2}}$$

= 7.23 i.e. Rs.7.23 lakh.

SUMMARY

- Correlation can be defined as the degree of association between two variables. The measure of correlation is called correlation coefficient and it ranges between 0 and 1. Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearson's Coefficient of Correlation is most widely used in practice.
- Simple correlation studies the association between two variables whereas Multiple correlation involves measurement of the degree of association between more than two variables, with one variable being dependent and others being independent. If we seek to discover just the influence of one independent variable on the dependent variable while taking into consideration the fact that the other variables are also operative, then we are measuring Partial correlation.
- In situations where correlation cannot be measured, rank correlation is used to measure the degree of agreement between two sets of ranks.

Chapter X

Regression Analysis

After reading this chapter, you will be conversant with:

- Meaning and Definition of Regression Analysis
- Application of Regression Analysis
- Difference between Correlation and Regression Analysis
- Regression Line
- Regression Equations
- Standard Error of Estimate
- Additional Illustrations

Introduction

From the previous chapter, it is clear that the two variables are closely related to each other. After establishing such relationship, we may be interested in predicting the value of one variable when the value of another variable is given. For example, if we know that supply and demand or sales and advertisement expenses are correlated, we can find out/predict the expected amount of supply/sales for a given demand/advertisement expenses or the demand/required advertisement amount for achieving the given supply/sales. Thus, estimation or prediction is possible through regression analysis because it reveals the average relationship between two variables.

MEANING AND DEFINITION OF REGRESSION ANALYSIS

The literal or dictionary meaning of the term *'regression' is 'the act of stepping back or returning to the average value'*. This term was first used by Sir Francis Galton an British Biometrician for studying the relationship between the heights of fathers and sons i.e. he estimated the extent to which the stature of the sons of tall parents reverts to the mean stature of the population. He published the result in the 19th century in a paper *'Regression towards Mediocrity in Hereditary Stature'*. He studied the relationship between the heights of about 1000 fathers and sons and revealed an interesting relationship:

1. The tall fathers will have tall sons and short fathers will have short sons.
2. But the average height of sons of group of tall fathers is less than that of the fathers and the average heights of the sons of a group of short fathers is more than that of the fathers.

Galton described this phenomenon, as Regression to Mediocrity i.e., the offspring's of abnormally tall or short parents tends to regress back to the average height of the population. The line that describes the tendency to regress back is called 'Regression Line' by Galton.

Today, the regression analysis is widely used in statistics and almost in all the sciences. It is widely and specially used in economics and business for studying the relationship between two or more variables that are closely related. Estimation or prediction are of paramount importance to an economist or businessman. In general sense, Regression Analysis means predicting or estimating the unknown value of one variable from the known value of another variable.

M.M.Blair *defines, "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data."*

According to ***Ya-Lun-Chou****, "Regression Analysis attempts to establish the nature of the relationship between variables – that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting."*

According to ***Morris Hamburg****, the term 'regression analysis' refers to "the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process."*

From the above definitions, it is clear that regression analysis is a statistical tool used for estimating the value of an unknown variable from the known value of another variable. In our day-to-day life, we come across various inter-related events such as crop yield depends on rainfall, cost of a product depends on production and advertising expenses, persons expenditure depends on his income, etc. The regression analysis for studying two variables at a particular time is termed Simple Linear regression analysis because only one variable is to be predicted or estimated. It is linear because the relationship between the two variables is assumed to be linear i.e. the equation of straight line is $Y = a + bX$. When the regression analysis is confined to studying the relationship between more than two variables at a time it is known as Multiple regression analysis.

A regression analysis will have two types of variables. The variable whose value we are predicting or estimating is called dependent or explained variable. The dependent variable is also known as regressed variable. The variable, which influences the value or is used for predicting the value of interested variable is called independent or explanatory variable. The independent variable is also known as regresser or predictor. The independent variable is denoted by X and the dependent variable by Y.

The relationship between the variables may be linear or non-linear. A linear relationship between variables exist when the regression curve is a straight line – Y = a + bX. The bivariate data when plotted on a graph, the points so obtained will concentrate round a curve, called the '*curve of regression*.' If the curve of regression is not a straight line, the regression is termed *Curved or non-linear regression.*

APPLICATION OF REGRESSION ANALYSIS

Regression analysis is widely used in all the scientific disciplines. It is a branch of statistical theory. It is used in economics for estimating or predicting the relationship between the economic variables. For example, if price and demand are the two closely related variables say X and Y, we can estimate the value of X for a given value of Y or the value of Y for a given value of X. The regression analysis is not only confined to business and economics but also extends to natural, physical and social sciences. The utility of regression analysis is explained as:

1. The regression analysis estimates the value of dependent variables from the value of independent variables through regression line. A regression line explains the relationship between the two variables say X and Y. The equation of this line is known as regression equation. The regression equation provides the estimated value of dependent variables when we insert the value of independent variable into the equation.
2. The regression analysis obtains measure of error involved while using the regression line for estimation. A standard error of estimate is calculated for this purpose. It measures the scatteredness of the observed value of dependent variables around the values estimated from the regression line. Good estimates of dependent variable can be made if the observations around the regression line are little scattered.
3. The coefficient of correlation can be obtained with the help of regression coefficients. We can also calculate the coefficient of determination, which is square of coefficient of correlation. The coefficient of determination measures the degree of correlation between the two variables.

Assumptions in Regression

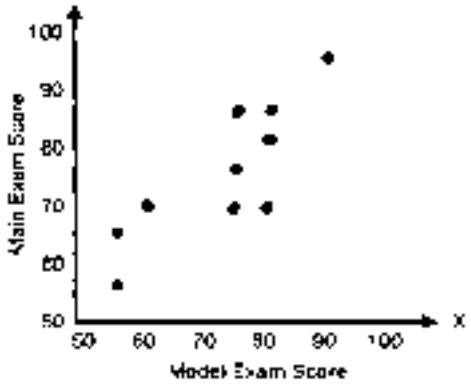
To understand the properties underlying the regression line, lets take an example.

Example

Students of the DBF program are required to give a model examination before they sit for the main exam of the program. The model exam is expected to give the students a chance of assessing their preparation under examination conditions. For a set of ten students, the scores in the model exam and the corresponding main exam are as follows:

Student Number	Model Exam Score (X)	Main Exam Score (Y)
1	55	65
2	90	95
3	75	75
4	80	70
5	75	85
6	60	70
7	80	80
8	55	55
9	80	85
10	75	70

Solution



From the scatter diagram, it appears that most students with high model exam scores have high main exam scores and students who scored low points in the model exam have also scored low points in the main exam. Thus, it appears that, the model exams are useful in predicting the students' scores for the main exam. So, if a student has scored 85 points in the model exam, is it possible to predict what will be his points in the main exam? By inspection of the figure, we might guess that this student will receive main exam points in the vicinity of 75 to 95.

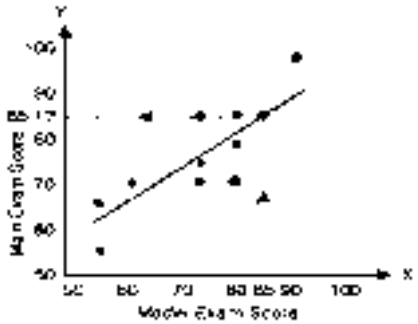
If we knew the model exam scores of all students along with their main exam scores, we would then have the population values. The mean and the variance of the population of the model exam would be μ_X and $\sigma^2_{x^2}$ respectively. The measurements for the main exam points are μ_Y and σ^2_y.

The assumptions in regression are:

1. The relationship between the variables X and Y is linear, which implies the formula $E(Y|X = x) = A + Bx$ at any given value of $X = x$.
2. At each X, the distribution of Y_x is normal, and the variances σ^2_X are equal. This implies that ε_X 's have the same variance, σ^2.
3. The Y-values are independent of each other.
4. No assumption is made regarding the distribution of X.

Since we do not have all of the students' course points and main exam points we must estimate the regression line $E(Y|X = x) = A + BX$.

The figure shows a line that has been constructed on the scatter diagram. Note that the line seems to be drawn through the collective mid-point of the plotted points. The term $\hat{Y}_X$ is the estimate of the true mean of Y's at any particular $X = x$.



DIFFERENCE BETWEEN CORRELATION AND REGRESSION ANALYSIS

For a given problem, it is difficult to decide which measure (i.e., whether correlation or regression) to be used, as both the measures give different information under different assumptions. The differences between correlation and regression are as follows:

1. The correlation coefficient measures the degree of co-variability between the two variables i.e., X and Y, whereas the regression analysis studies the nature of relationship that exists between them i.e., predicting the value of one variable on the basis of known value of other variable.
2. In correlation analysis, we study only the intensity and direction of the relationship between the two variables and cannot point out that one variable is cause and other is the effect. Whereas in regression analysis, one variable is taken as dependent variable and other as independent variable, which facilitates the study of cause and effect relationship among the variables.
3. Coefficient of correlation is independent of change of scale and origin. Whereas the coefficient of regression is independent of change of origin and not scale.
4. Sometimes, the correlation between the two variables may be due to chance and has no practical relevance. However, there is no such chance in regression analysis and has practical relevance.
5. In correlation r_{xy} and r_{yx} are symmetric i.e., it is immaterial which is the dependent and independent variable. Whereas in regression analysis, the coefficient b_{xy} and b_{yx} are not symmetric i.e., it makes difference as to, which variable is dependent and which is independent.

REGRESSION LINE

Regression line is a line that gives the best estimation of one variable for any given value of other variable. If we are having two variables say X and Y, we will have two regression lines i.e., regression of X on Y and the regression of Y on X.

Line of regression of Y on X is the line which gives the estimated value of Y for any specified value of X.

Similarly, the line of regression of X on Y is the line which gives the estimated value of X for any specified value of Y.

These two regression lines will coincide when there is either a perfect positive or perfect negative correlation between the two variables. The degree of correlation is less when the two regression lines are far from each other and the degree of correlation is higher when the regression lines are near to each other. The regression lines are parallel to OX and OY axis and the r will be zero, if the variables are independent. The regression line cut each other at a point from where we get the mean value of X if a perpendicular is drawn on the X-axis and we can also obtain the mean value of Y if a horizontal line is drawn on the Y-axis.

The regression lines or line of best fit are drawn in accordance with the Principle of Least Square. The principle of least square consists in minimizing the sum of the squares of the residuals or errors of estimates, i.e., the deviations between the given observed values of the variable and their corresponding estimated values as given by the line of best fit. In other words,

> The regression line should be drawn on the scatter diagram in such a way that when the squared values of the vertical distance from each plotted point to the line are added, the total amount will be the smallest possible amount. This criterion is called the Method of Least Squares.

It assumes that the line of best fit gives the minimum sum of squares of the deviations. For minimizing this, we will have two regression lines – vertical parallel to Y-axis and horizontal parallel to X-axis. The line of regression of Yon X minimizes the sum of square of deviations parallel to Y-axis and the line of regression of X on Y minimizes the sum of squares of deviations parallel to X-axis.

REGRESSION EQUATIONS

The algebraic expression of the regression lines is known as Regression Equation or Estimating equations. As we have two regression lines, we have two regression equations – regression equation of Y on X and regression equation of X on Y. The regression equation of x on y describes the variations in the values of X for a given variation/change in Y and the regression equation of y on x explains the variation in the values of Y for a given change in X.

Regression Equation of Y on X

The regression equation of Y on X is expressed as follows:

$$Y = a + bX$$

In the above equation, X and Y are the two variables. Y is a dependent variable whose value depends on X, and X is an independent variable from which we can compute the value of Y. 'a' and 'b' in the equations are numerical constants as their values does not change for a given straight line. 'a' is Y-intercept because at this point, the regression line crosses the Y-axis. 'b' is the slope of line of regression of y on x. It is called the coefficient of regression of y on x which measures the change in the value of dependent variable Y for a unit change in the independent variable X.

The regression line will be completed, once the values of a and b are determined. These values are determined through the method of least square. The method of least square states that through the plotted points, a line is to be drawn in such a manner that the sum of the square of the deviations of actual dependent variable (Y) from the computed dependent variable (Y) is minimum. Such a line is known as line of 'best fit'. The straight line has the following characteristics:

i. The sum of the deviations obtained from the line, $\sum(Y - Y_c)^2$, is less than they would be from any other straight line.

ii. The deviations above the line are equal to the deviations below the line i.e. $\sum(Y - Y_c) = 0$

iii. The straight line passes through the overall mean of the data i.e., $\overline{X}\overline{Y}$

iv. The least square line is a best estimate of the population regression line when the sample is taken from the large population.

So the regression line of y on x is expressed as follows:

$$Y_c = a + bX$$

The values of constants a and b are determined by solving the following equations Simultaneously:

$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

where $\sum X$, $\sum XY$, and $\sum X^2$ represent the sum obtained from the observed pairs of values of variables X and Y, and N indicates the number of observed pair of values.

Regression Equation of X on Y

The regression equation of Y on X is expressed as follows:

$X_c = a + bY$

The values of constants a and b are determined by solving the following equations Simultaneously:

$\sum X = Na + b\sum Y$

$\sum XY = a\sum Y + b\sum Y^2$

Illustration 1

The following table gives the information relating to prices and demand for a product.

Prices	1	2	3	4
Demand	8	6	4	2

Find the two regression equations for the above data.

Solution

Calculation of Regression Equation

Price (X)	Demand (Y)	X^2	Y^2	XY
1	8	1	64	8
2	6	4	36	12
3	4	9	16	12
4	2	16	4	8
$\sum X = 10$	$\sum Y = 20$	$\sum X^2 = 30$	$\sum Y^2 = 120$	$\sum XY = 40$

Regression equation of X *on* Y

$X_c = a + bY$

The two normal equations are,

$\sum X = Na + b\sum Y$

$\sum XY = a\sum Y + b\sum Y^2$

Substituting the values

$10 = 4a + b20$ … (1)

$40 = 20a + b120$ … (2)

Multiply the equation (1) by 5

$50 = 20a + b100$ … (3)

$40 = 20a + b120$ … (4)

deducting the equation (4) from (3) $10 = -20b$

$b = -0.50$

Substitute the value of b in equation (1)

$10 = 4a + 20\ (-0.50)$ or $4a = 10 + 10 = 20$ or $a = 5$

Put the calculated values of a and b in the equation, the regression equation of X on Y is

$X = 5 - 0.50Y$

Regression line of Y on X

$Y_c = a + bX$

The two normal equations are:

$\sum Y = Na + b\sum X$

$\sum XY = a\sum X + b\sum X^2$

$20 = 4a + 10\ b$ … (1)

$40 = 10\ a + 30\ b$ …(2)

Multiply the equation (1) by 2.5

$50 = 10a + 25\ b$ … (3)

$40 = 10a + 30\ b$ … (4)

From equation (3) and (4) $-5\ b = 10$ 0r $b = -2$

Substitute the values of b in the (1) equation ;

$20 = 4a + 10\ (-2)$ or $4a = 20 + 20 = 40$ or $a = 10$

Put the calculated value of a and b in the equation, the regression equation of Y on X is

$Y = 10 - 2\ X.$

Deviations Taken from Arithmetic Mean X and Y

The method for finding out the regression equation of X on Y and Y on X discussed above is tedious. These calculations can be simplified, if we take deviation of X and Y from their respective means instead of taking the actual values. When the deviations are taken from the arithmetic mean of X and Y, the two regression equations are expressed as follows:

Regression Equation of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

Where,

$\bar{X}$ is the arithmetic mean of series X;

$\bar{Y}$ is the arithmetic mean of series Y;

$r\frac{\sigma_x}{\sigma_y}$ is the regression coefficient of X on Y which can be denoted by the symbol as b_{xy}. As said earlier, it measures the change in the value of X due to the unit change in the value of Y. when the deviations are taken from the mean of X and Y, the coefficient of regression of X and Y can also be obtained by the following formula:

$$b_{xy} \text{ or } r\frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2} \text{ or } \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2}$$

The value of regression coefficient can also be determined by calculating $\sum xy$ and $\sum y^2$ instead of the correlation coefficient(r), σ_x and σ_y.

Regression Equation of Y on X

$$(Y - \bar{Y}) = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

The regression coefficient of Y on X $r\frac{\sigma_y}{\sigma_x}$ is denoted by b_{yx}, which measures the change in the value of y corresponding to the unit change in the value of X. When we take deviations from the actual mean, the regression coefficient of Y on X is calculated as follows:

$$r\frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2} \text{ or } \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

From the two regression coefficients, we can calculate the coefficient of correlation. The coefficient of correlation is the square-root of the product of two regression coefficients. The correlation coefficient is expressed as follows:

$$r = \sqrt{b_{xy} \times b_{yx}}$$

Where

$$b_{xy} = r\frac{\sigma_x}{\sigma y} \text{ and } b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

Important points relating to regression coefficient:

1. The two regression coefficients will have the same sign i.e., both will be either positive or negative.
2. Usually, the value of coefficient of correlation does not exceed one i.e., it lies between $\pm$ 1.
3. The regression coefficient and coefficient of correlation will have same sign i.e., if the regression coefficient is negative, the correlation coefficient will be negative.
4. Regression coefficients are independent of change in origin but not of scale.

Illustration 2

Lets take the data of illustration 1 and calculate the regression equations taking deviations of items from the mean of X and Y series.

Solution

Calculation of Regression Equations

X	$X - \bar{X} = x$	x^2	Y	$Y - \bar{Y} = y$	y^2	xy
1	–1.5	2.25	8	3	9	–4.5
2	–.5	0.25	6	1	1	–0.5
3	.5	0.25	4	–1	1	–0.5
4	1.5	2.25	2	–3	9	–4.5
$\sum X = 10$		$\sum x^2 = 5$	$\sum Y = 20$		$\sum y^2 = 20$	$\sum xy = -10$

Regression equation of X on Y

$$(X - \bar{X}) = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$\bar{X} = \frac{\sum X}{N} = \frac{10}{4} = 2.5\,;\ \bar{Y} = \frac{\sum Y}{N} = \frac{20}{4} = 5\,;\quad r\frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2} = \frac{-10}{20} = -0.50$$

$$(X - 2.5) = -0.50\,(Y - 5)$$

$$(X - 2.5) = -0.50Y + 2.5$$

$$X = -0.50Y + 2.5 + 2.5$$

$$X = -0.50Y + 5.0 \text{ or } X = 5 - 0.50Y$$

Regression equation of Y on X

$$(Y - \bar{Y}) = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2} = \frac{-10}{5} = -2$$

$$(Y - 5) = -2\,(X - 2.5)$$

$$(Y - 5) = -2X + 5$$

$$Y = -2X + 5 + 5$$

$$Y = -2X + 10 \text{ or } Y = 10 - 2X$$

This method simplifies the calculation when compared to the normal equations used earlier.

DEVIATIONS TAKEN FROM ASSUMED MEAN

When the actual mean of X and Y are in fractions, the calculation can be simplified by taking the deviations from the assumed mean. When the deviations are taken from assumed mean the procedure for computation of regression equations remains the same but the only difference is that the deviations are obtained from assumed mean instead of actual mean. The regression equations are as follows:

Regression equation of X on Y

$$(X - \bar{X}) = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

but the value of $r\frac{\sigma_x}{\sigma_y} = \frac{N\sum d_x d_y - (\sum d_x \times \sum d_y)}{N\sum d_y{}^2 - (\sum d_y)^2}$

$$d_x = (X - A)\,;\, d_y = (Y - A)$$

Regression equation of Y on X

$$(Y - \bar{Y}) = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{N\sum d_x d_y - (\sum d_x \times \sum d_y)}{N\sum d_x{}^2 - (\sum d_x)^2}$$

Illustration 3

From the data given in Illustration 1, compute the regression equation when the mean is assumed to be 3 and 6.

Solution

X	$X-3=(d_x)$	d_x^2	Y	$(Y-6)=d_y$	d_y^2	$d_x\ d_y$
1	–2	4	8	2	4	– 4
2	–1	1	6	0	0	0
3	0	0	4	– 2	4	0
4	1	1	2	– 4	16	– 4
$\sum X = 10$	$\sum d_x = -2$	$\sum d_x^2 = 6$	$\sum Y = 20$	$\sum dy = -4$	$\sum d_y^2 = 24$	$\sum d_x\ d_y = -8$

Regression Equation of X on Y

$$(X - \bar{X}) = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y}) \qquad \bar{X} = \frac{\sum X}{N} = \frac{10}{4} = 2.5\ ;\ \bar{Y} = \frac{\sum Y}{N} = \frac{20}{4} = 5\ ;$$

$$r\frac{\sigma_x}{\sigma_y} = \frac{N\sum d_x d_y - (\sum d_x \times \sum d_y)}{N\sum d_y^2 - (\sum d_y)^2}$$

$$= \frac{4(-8) - (-2)(-4)}{4(24) - (-4)^2}$$

$$\frac{-32 - 8}{96 - 16} = \frac{-40}{80} = -0.50$$

$$(X - 2.5) = -0.50\ (Y - 5)$$

$$(X - 2.5) = -0.50Y + 2.5$$

$$X = -0.50Y + 2.5 + 2.5$$

$$X = -0.50Y + 5.0 \quad \text{or} \quad X = 5 - 0.50Y$$

Regression equation of Y on X

$$(Y - \bar{Y}) = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$r\frac{\sigma_x}{\sigma_y} = \frac{N\sum d_x d_y - (\sum d_x \times \sum d_y)}{N\sum d_x^2 - (\sum d_x)^2}$$

$$\frac{4(-8) - (-2)(-4)}{4(6) - (-2)^2}$$

$$\frac{-32 - 8}{24 - 4} = \frac{-40}{20} = -2$$

$$(Y - 5) = -2\ (X - 2.5)$$

$$(Y - 5) = -2X + 5$$

$$Y = -2X + 5 + 5$$

$$Y = -2X + 10 \text{ or } Y = 10 - 2X$$

From the above calculations, it is clear that the answer remains the same whether deviations are taken from actual mean or assumed mean.

Graphing the Regression Lines

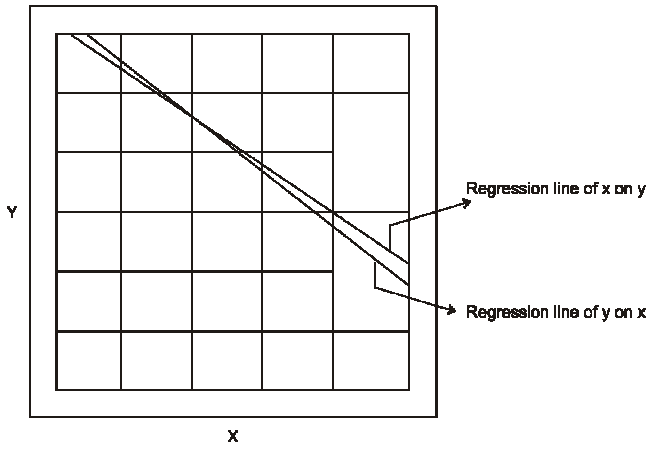
The regression lines can be easily plotted on the graph once they are calculated. Following procedure is followed while plotting the lines on graphs:

- For the unknown variables of the equation choose any two values,
- Calculate the other variable from it,
- Then plot the pairs of value and draw a straight line through the plotted points.

Illustration 4

Show the regression equation of illustration no.3 graphically.

Solution



Regression line of X on Y [5 – 0.50 Y]

Let Y = 3, X = 5 – 0.50 (3) = 5 –1.50 = 3.50

Let Y = 5, X = 5 – 0.50 (5) = 5 – 2.50 = 2.50

Regression line of Y on X [10 – 2X]

Let X = 1, Y = 10 – 2(1) = 10 – 2 = 8

Let X = 5 , Y = 10 – 2(5) = 10 –10 = 0.

STANDARD ERROR OF ESTIMATE

The measure of reliability of the estimating equation that we have developed is given by standard error of estimate. The standard error of estimate represented by s_e is similar to the standard deviation as both are measures of dispersion. The standard error measures the variability, or scatterness, of the observed values around the regression line.

The standard error of the estimate for a regression equation is given by,

$$s_e = \frac{\sqrt{\Sigma \left(Y - \hat{Y}\right)^2}}{n-2}$$

Where,

Y = values of the dependent variable.

$\hat{Y}$ = estimated values from the estimating equation that correspond to each Y value.

n = number of data points used to fit the regression line.

In the above equation, you can observe that the sum of the squared deviations is divided by n – 2 and not by n. This is because we have lost 2 degrees of freedom in estimating the regression line as there are two unknown constants. We used the sample to compute a and b.

Equation for calculating standard error of estimate is also given by formula,

$$s_e = \frac{\sqrt{\sum Y^2 - a\sum Y - b\sum XY}}{n-2}$$

Illustration 5

Consider the Illustration of a Pharmaceutical Firm. Find the standard error.

(X)	(Y)	$\hat{Y}$	$(Y - \hat{Y})$	$\left(Y - \hat{Y}\right)^2$
5	31	30	1	1
11	40	42	–2	4
4	30	28	2	4
5	34	30	4	16
3	25	26	–1	1
2	20	24	–4	16
Total 30	180	180	0	42

Solution

Standard error of estimate $s_e = \frac{\sqrt{\sum \left(Y - \hat{Y}\right)^2}}{n-2}$

Substituting the values in the above equation,

$s_e = \sqrt{42/4} = 3.24$

Standard error = 3.24 crore.

Similar to standard deviation, the larger the standard error of estimate, the greater the dispersion or scattering of points around the regression line. If standard error of estimate is zero, the estimating equation is a perfect estimator of the dependent variable. Thus, all the points would be directly on the regression line, and no points would be scattered around it.

$s_{\hat{e}}^2$ is an unbiased estimator of σ^2 the variance of the ε_x's

Prediction Interval

We would like to construct a prediction interval around $\hat{Y}$ which would contain the actual Y.

If $n \geq 30$, $\hat{Y} \pm Zs_e$ would be the interval, where Z is the appropriate Standard Normal Value.

If $n < 30$, $\hat{Y} \pm ts_e$ would be the interval, where 't' is the appropriate 't' value (with n – 2 degrees of freedom).

In the case of the pharmaceutical firm, at X = Rs.1 crore, the 90% prediction interval is 22 ± 2.132 x 3.24.

Coefficient of Determination

Given the paired observations (X, Y), we can estimate the value of Y in two ways.

i. $\overline{Y}$

ii. $\hat{Y} = a + bx$.

We may then compute $\sum\left(Y-\overline{Y}\right)^2$ and $\sum\left(\hat{Y}-\overline{Y}\right)^2$. The former is defined as the total variation in Y and the latter is called variation explained by the model.

Define $R^2 = \dfrac{\sum\left(\hat{Y}-\overline{Y}\right)^2}{\sum\left(Y-\overline{Y}\right)^2}$

It can be shown that $\sqrt{R^2}$ = r, the coefficient of correlation.

Therefore $0 \le R^2 \le 1$, and greater the R^2, better the fit.

We define the variables,

TSS (Total Sum of Squares) = $\sum\left(Y-\overline{Y}\right)^2$

RSS (Regression Sum of Squares) = $\sum\left(\hat{Y}-\overline{Y}\right)^2$

ESS (Error Sum of Squares) = $\sum\left(Y-\hat{Y}\right)^2$

We can show that TSS = RSS + ESS

Coefficient of determination is also calculated by $R^2 = \dfrac{a\sum Y + b\sum XY - n\overline{Y}^2}{\sum Y^2 - n\overline{Y}^2}$

Illustration 6

For the pharmaceutical firm problem, let us understand the geometric concept behind RSS, ESS, TSS and R^2.

Solution

For a whole set of values of y, consider the table given below:

Y	$\hat{Y}$	$\left(Y-\overline{Y}\right)^2$	$\left(\hat{Y}-\overline{Y}\right)^2$	$\left(Y-\hat{Y}\right)^2$
31	30	1	0	1
40	42	100	144	4
30	28	0	4	4
34	30	16	0	16
25	26	25	16	1
20	24	100	36	16
Total 180	180	242	200	42

In this case $\sum\left(Y-\overline{Y}\right)^2 = \sum\left(\hat{Y}-\overline{Y}\right)^2 + \sum\left(Y-\hat{Y}\right)^2$.

To continue with our discussion, we have $R^2 = \frac{RSS}{TSS} = 0.83$ and note that R^2 measures how good the fit is. In the extreme case when there is no fit, RSS = 0 and $R^2 = 0$. In the other extreme case when there is a perfect fit (that is, all the observations fall on a straight line), RSS = TSS and $R^2 = 1$ and r is either 1 or –1.

ADDITIONAL ILLUSTRATIONS

Illustration 1

Calculate the regression equation for the following data relating to model exam score (X) and main exam score (Y)

X	55	90	75	80	75	60	80	55	80	75
Y	65	95	75	70	85	70	80	55	85	70

Solution

Student	Y	Y^2	X	X.Y	X^2
1	65	4,225	55	3,575	3,025
2	95	9,025	90	8,550	8,100
3	75	5,625	75	5,625	5,625
4	70	4,900	80	5,600	6,400
5	85	7,225	75	6,375	5,625
6	70	4,900	60	4,200	3,600
7	80	6,400	80	6,400	6,400
8	55	3,025	55	3,025	3,025
9	85	7,225	80	6,800	6,400
10	70	4,900	75	5,250	5,625
Total	750	57,450	725	55,400	53,825

The coefficient b is,

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$$= \frac{10(55,400) - (725)(750)}{10(53,825) - (725)^2}$$

$$= \frac{554,000 - 543,750}{538,250 - 525,625} = \frac{10,250}{12,625} = 0.81$$

The coefficient a is,

$$a = \bar{Y} - b\bar{X}$$

$$= \frac{\sum Y}{n} - b\frac{\sum X}{n}$$

$$= \frac{750}{10} - (0.81)\frac{725}{10}$$

$$= 75 - 58.73 = 16.27$$

Therefore, the equation for the regression line is $\hat{Y}_X = 16.27 + 0.81X$.

From the slope of this regression (+ 0.81) it can be stated that for every unit increase or additional point on the model exam, it is estimated that the average main exam points will increase by 0.81 points.

We can use this equation to find the point estimate of the average main exam marks of all students who have any particular model exam marks X. For the average of all students with a score of 85 on the model exam, the predicted main exam points, based on this sample of ten X,Y pairs will be,

$$\hat{Y}_X = 16.27 + 0.81(85) = 16.27 + 68.85 = 85.12$$

Illustration 2

A pharmaceutical firm would like to establish the relationship between the annual expenditure on research and development and the annual profits. Data for the past few years is given below:

	Research and Development Expenditure	Annual Profits
Year	Rs. in crore (X)	Rs. in crore (Y)
2001	5	31
2000	11	40
1999	4	30
1998	5	34
1997	3	25
1996	2	20

Solution

For a given value of x, y is a random variable with a mean and standard deviation. Note that at x = 5 Crore, y has assumed the values Rs.31 crore and Rs.34 crore in different years.

The intercept 'a' and slope 'b' can be determined by setting up the table.

X	Y	XY	X^2
5	31	155	25
11	40	440	121
4	30	120	16
5	34	170	25
3	25	75	9
2	20	40	4
Total 30	180	1000	200

$$\bar{X} = \frac{\sum X}{n} = \frac{30}{6} = 5, \quad \bar{Y} = \frac{\sum Y}{n} = \frac{180}{6} = 30$$

$$\sum XY = 1000, \qquad \sum X^2 = 200$$

$$\bar{X}^2 = (5 \times 5) = 25$$

We know that,

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{1000 - (6)(5)(30)}{200 - (6)(25)} = \frac{100}{50} = 2$$

$$a = \bar{Y} - b\bar{X} = 30 - (2)(5) = 20$$

The estimating equation is,

$$\hat{Y} = 20 + 2X$$

Where $\hat{Y}$ is the estimate of annual profits when R & D expenditure is X.

Illustration 3

Consider the following data on the number of hours which 10 persons studied for a statistics test and their scores on the test.

Hours Studied (X)	4	9	10	14	4	7	12	22	1	17
Test Score (Y)	31	58	65	73	37	44	60	91	21	84

a. Fit a regression equation of the form Y = a + bX.

b. Calculate the standard error of estimate.

c. Calculate the coefficient of determination. What does it convey?

Solution

Y	Y^2	X	X.Y	X^2
31	961	4	124	16
58	3364	9	522	81
65	4225	10	650	100
73	5329	14	1022	196
37	1369	4	148	16
44	1936	7	308	49
60	3600	12	720	144
91	8281	22	2002	484
21	441	1	21	1
84	7056	17	1428	289
564	36562	100	6945	1376

We need to fit a straight line of the form

$$Y = a + bX$$

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$$= \frac{10(6945) - (100)(564)}{10(1376) - (100)^2} = 3.4707$$

$$a = \bar{Y} - b\bar{X}$$

$$= 564 / 10 - 3.4707\ (100 / 10) = 21.693$$

= Standard error of estimation

$$s_e = \frac{\sqrt{\sum Y^2 - a\sum Y - b\sum XY}}{n-2}$$

$$= \sqrt{\frac{36562 - (21.693)(564) - (3.4707)(6945)}{12-2}}$$

$$= 4,724$$

The coefficient of determination, r^2 is given by

$$r^2 = \frac{a\sum Y + b\sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}$$

$$= \frac{(21.693)(564) + (3.4707)(6945) - 10(56.4)^2}{36562 - 10(56.4)^2}$$

$$= 0.9530$$

Therefore, 95.3% of the variations in the test scores is explained by the regression line. We could also say that 95.3% of the variations in the dependent variable, test score is explained by the independent variable number of hours studied.

Illustration 4

Consider the following data relating to 9 salesman of a company in an intelligence test and their sales for the week are given.

Salesman	A	B	C	D	E	F	G	H	I
Intelligent Scores	40	50	40	50	70	40	70	30	60
Sales Rs. ('000)	20	50	30	40	50	20	60	40	50

a. Calculate the regression equation of sales on salesman's intelligence test score.

b. If the salesman's intelligence score is 55, what would be his expected weekly sales?

Solution

Calculation of Regression Equations

X	X – 50 = x	x^2	Y	Y– 40 = y	y^2	xy
40	–10	100	20	–20	400	+200
50	0	0	50	+10	100	0
40	–10	100	30	–10	100	+100
50	0	0	40	0	0	0
70	+20	400	50	+10	100	+200
40	–10	100	20	–20	400	+200
70	+20	400	60	+20	400	+400
30	–20	400	40	0	0	0
60	+10	100	50	+10	100	+100
$\sum X = 450$		$\sum x^2 = 1600$	$\sum Y=360$		$\sum y^2 =1600$	$\sum xy=1200$

Regression equation of Y on X

$$(Y - \bar{Y}) = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2} = \frac{1200}{1600} = 0.75 \quad \bar{X} = \frac{\sum X}{N} = \frac{450}{9} = 50; \quad \bar{Y} = \frac{\sum Y}{N} = \frac{360}{9} = 40$$

$(Y – 50) = 0.75(X – 60)$

$(Y – 5\ 0) = 0.75X – 45$

$Y = 5 + 0.75X$

Expected sales for a week when the intelligence score of a saleman is 55.

$Y = 5 + 0.75X$ (where $X = 55$)

$Y = 5 + 41.25$ (0.75 x 55)

$= 46.25$

Illustration 5

In a correlation study the following values are obtained:

	X	Y
Mean	55	57
Standard deviation	2.5	3.5
Coefficient of correlation	0.8	

Compute the two regression equation for the above data.

Solution

Regression equation of Y on X

$$(Y - \bar{Y}) = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$(Y - 57) = 0.80\frac{3.5}{2.5}(X - 55)$$

$Y - 57 = 1.12\,(X - 55)$

$Y - 57 = 1.12X - 61.6$

$Y = 1.12X - 61.6 + 57$

$Y = 1.12X - 4.6$ or $Y = -4.6 + 1.12X$

Regression equation of X on Y

$$(X - \bar{X}) = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$(X - 55) = 0.80\frac{2.5}{3.5}(Y - 57)$$

$X - 55 = 0.5714\,(Y - 57)$

$X = 0.5714Y - 32.5698 + 55$

$X = 0.5714Y + 22.4302$ or $X = 22.43 + 0.5714Y$.

Illustration 6

For certain X and Y series which are correlated, the two regression lines are:

$4X - 5Y + 33 = 0$

$20X - 9\,Y = 107$

Find i. The mean value of X and Y.

ii. Coefficient of correlation between X and Y.

iii. Standard deviation of Y when the variance of X = 9.

Solution

i. Calculation of Mean value of X and Y

$4X - 5Y = -33$

$20X - 9\,Y = 107$

Multiply the (1) equation by 5

$20X - 25Y = -165$

$20X - 9Y = 107$

$- \quad + \quad -$

$16Y = -272$

$Y = 17$ or $\bar{Y} = 17$

By substituting the value of Y in the (1) equation, we can obtain the value of X

$4X - 5Y = -33$

$4X - 5(17) = -33$

$4X = -33 + 85$

$X = 52/4 = 13$ or $\bar{X} = 13$

ii. For calculating the correlation coefficient, we make certain assumption. Lets take (1) equation:

$4X = -33 + 5Y$

$$X = -\frac{33}{4} + \frac{5}{4}y$$

$$\text{or } b_{xy} = \frac{5}{4} = 1.25$$

From (2) equation, we can calculate b_{yx}

$$-9Y = 107 - 20X$$

$$Y = -\frac{107}{9} + \frac{20}{9}X$$

$$\text{or } b_{yx} = \frac{20}{9} = 2.22$$

Since, the value of regression equation is more than 1, our assumption is wrong. Hence, the equation (1) is the equation Y on X.

From equation (1) $= -5Y = -4X - 33$

$$Y = \frac{4}{5}X + 3.3 \text{ or } b_{yx} = \frac{4}{5}$$

$$b_{xy} = \frac{9}{20} = 0.45$$

From equation (2)

$$r = \sqrt{\frac{4}{5} \times \frac{9}{20}} = \sqrt{0.36} = 0.6$$

$$\sigma_x = \sqrt{9} = 3$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$0.45 = 0.6\frac{3}{\sigma_y}$$

$$0.45\sigma_y = 1.8$$

$$\sigma_y = \frac{1.8}{0.45} = 4$$

Illustration 7

Northern Transport Company is planning to start an overnight luxury bus service between Delhi and Jaipur from March, 2002. A survey was conducted to assess the number of passengers available per day at different ticket prices. The results are given below:

Price Per Ticket (Rs.)	250	275	300	325	350
Number of Passengers Per Day	110	90	80	70	50

If equation of straight line is formed to predict number of passangers per day what will be the slope of line?

Solution

Let the following notations be used:

Price per ticket: x

Number of passengers per day: y

Sl. No.	Ticket Prices (x)	Number of passengers per day (y)	x^2	xy
1	250	110	62,500	27,500
2	275	90	75,625	24,750
3	300	80	90,000	24,000
4	325	70	105,625	22,750
5	350	50	122,500	17,500
	$\Sigma x = 1,500$	$\Sigma y = 400$	$\Sigma x^2 =$ 4,56,250	$\Sigma xy = 1,16,500$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{400}{5} = 80, \quad \bar{x} = \frac{1500}{5} = 300$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{5 \times 1,16,500 - 1,500 \times 400}{5 \times 4,56,250 - (1,500)^2} = \frac{-17,500}{31,250} = -0.56$$

$$a = \bar{y} - b\bar{x} = 80 - (-0.56) \times 300 = 248$$

$\therefore$ The regression equation is $\hat{y} = 248 - 0.56x$

Illustration 8

A simple regression relationship was developed between the variables X and Y, with X as the independent variable. The estimated value of Y when X = 2, is 7.0; and the estimated value of Y when X = 8, is 11.8.

$\Sigma Y^2 = 791.31$ $\Sigma XY = 452.7$ $\Sigma Y = 78.1$ $n = 8$

Calculate the standard error of estimate for the regression line.

Solution

$$\hat{Y} = a + bX$$

Given: $7 = a + b(2)$ (For X = 2)

or $7 = a + 2b$ … (A)

$11.8 = a + b(8)$ (For X = 8)

or $11.8 = a + 8b$ … (B)

$$\therefore b = \frac{(11.8 - 7)}{(8 - 2)} = 0.80$$

Putting b = 0.80 in (A) $7 = a + 2(0.80)$

or $a = 7 - 1.6 = 5.4$

$\therefore$ The regression equation is

$$\hat{Y} = 5.4 + 0.80X$$

Standard error of estimate;

$$S_e = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n - 2}}$$

$\Sigma Y^2 =$ 791.31 $\Sigma XY = 452.7$

$\Sigma Y =$ 78.1 $n = 8$

$$\therefore S_e = \sqrt{\frac{791.31 - (5.4 \times 78.1) - (0.8 \times 452.7)}{8 - 2}}$$

$$= 1.1113$$

SUMMARY

- Simple linear regression computes the model that best fits the relationship indicated by coefficient of correlation. Before performing correlation and regression analysis, the cause and effect relationship between the variables should be understood. The assumptions in regression are that the relationship between the distributions X and Y is linear; at each X, distribution of Y is normal and the variance is equal; the Y values are independent of each other.

- The regression line should be drawn on a scatter diagram in such a way that when the squared values of the vertical distance from each plotted point is added, the total amount will be the smallest amount. This criterion is called the method of least squares. Using this principle, we obtain the normal equations, which when solved gives us the regression constants. Thus, if a causal relationship is established, we can fit a regression line, which can be used to predict the value of the dependent variable for a given value of the independent variable.
- The serious limitation in using regression analysis is that past trends are used to estimate future trends. The correlation and regression analysis are applied in the field of finance to find the risk of a portfolio in Cost-Volume-Profit analysis, in time series and demand forecasting.

Chapter XI

Index Numbers

After reading this chapter, you will be conversant with:

- The Concept of Index Numbers
- Types of Index Numbers
- Methods of Constructing Index Numbers
- Aggregates Method
- Average of Relatives Method
- Value Index Numbers
- Tests for Consistency
- Consumer Price Index Number
- Additional Illustrations

Introduction

An index number is a statistical measure designed to show changes in a variable or a group or related variables with respect to time, geographic location or other characteristics such as income, profession, etc. Index number is calculated as a ratio of the current value to a base value and expressed as a percentage. It must be clearly understood that the index number for the base year is always 100. An index number is commonly referred to as an index.

Today, Index numbers are one of the most widely used statistical indicators. Generally, they are used to indicate the state of the economy, index numbers are aptly called 'barometers of economic activity'. Index numbers are used for comparing production, sales or changes in exports or imports over a certain period of time. They is also used for sensex purposes at stock exchanges and for comparing changes in the economy through inflation, GDP, etc. The role played by index numbers in the Indian trade and industry is impossible to ignore. It is a very well-known fact that the wage contracts of workers in our country are tied to the cost of living index numbers.

THE CONCEPT OF INDEX NUMBERS

An index number is an average with a difference. An index number is used for the purpose of comparison in cases, where the series being compared could be expressed in different units, i.e., manufactured products index (a part of the wholesale price index) is constructed using items like Dairy Products, Sugar, Edible Oils, Tea and Coffee, etc. These items naturally are expressed in different units like sugar in kgs, milk in liters, etc. The index number is obtained as a result of an average of all these items, which are expressed in different units. On the other hand, average is a single figure representing a group expressed in the same units.

Index numbers essentially capture the changes in the group of related variables over a period of time. For example, if the index of industrial production is 215.1 in 2007-08 (base year 1999-00), it means that the industrial production in that year was up by 2.15 times compared to 1999-00. However, it mean that the net increase in the index reflects an equivalent increase in industrial production in all sectors of the industry. Some sectors may have increased their production more than 2.15 times while other sectors may have increased their production only marginally.

Definitions

Various statisticians have defined an index number in different forms. Some of them are studied below:

According to Spiegel "an index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographic location or other characteristics."

A.M. Tuttle defined index number as "a single ratio which measures to combine change of several variables between two different times, places or situations."

L.J. Kaplass stated that "an index number is a statistical measure of fluctuations in a variable arranged in the form of a series and a base period for making comparisions."

Thus, by above definitions we can summarize that "an index number is a ratio or an average of ratios expressed as a percentage. Two or more time periods are involved, one of which is the base time period. The value at the base time period

serves as the standard point of comparison". It is mathematically expressed in the following manner:

$$\text{Index Number} = \frac{\text{Current year Index}}{\text{Base year Index}} \times 100$$

Current Year Refers to the year for which the comparisons are sought and is denoted by the subscript '1'. E.g., if it is price index, it is denoted by P_1.

Base year refers to the year on the basis of which the comparisons are made and is denoted by the subscript '0'. E.g., if it is price index, it is denoted by P_0. The base year should be free from all kinds of abnormalities like flood, war, earthquake, fluctuation in the economic conditions etc.

Characteristics of Index Numbers

- Index number is a special type of average, which measures the change of variables between two different times, places and presents the average in percentages.
- Index numbers are expressed in percentages to show the extent of relative changes, but the symbol, % is not used to express it.
- Index numbers study the effect of changes in magnitude which cannot be measured directly. Examples of such phenomenon are price level, cost of living etc.
- Index numbers measure the relative change of differences from time to time or place to place. Thus, they facilitate comparison.

Uses of Index Numbers

1. **Establishes Trends**

   Index numbers when analyzed reveal a general trend of the phenomenon under study. For example: It helps in the study of variation in inflation, incomes etc., in different periods.

2. **Helps in Policy-Making**

   Index numbers guide policy-making. For example: it is widely known that the dearness allowance paid to the employees is linked to the cost of living index, generally, the consumer price index. From time to time, it is the cost of living, index which forms the basis of many wage agreements between the employees union and the employer.

3. **Determines Purchasing Power of the Rupee**

   Usually, index numbers are used to determine the purchasing power of the rupee. Suppose, the consumer price index for urban employees increased from 100 in 1996 to 202 in 2006, the real purchasing power of the rupee can be found out as follows:

   $$\frac{100}{202} = 0.495$$

   It indicates that if rupee was worth 100 paise in 1996, its purchasing power is 49.5 paise in 2006.

4. **Deflates Time Series Data**

   Index numbers play a vital role in adjusting the original data to reflect reality. For example, nominal income (income at current prices) can be transformed into real income (reflecting the actual purchasing power) by using income deflators. Similarly, assume that the industrial production is represented in

value terms as a product of volume of production and price. If the subsequent year's industrial production were to be higher by 20% in value, the increase might not be as a result of increase in the volume of production as one would have it, but because of increase in the price. The inflation which has caused the increase in the series can be eliminated by the usage of an appropriate price index and thus making the series real as presented in the Box.

Box 1: Construction of Real Wage Indices

Year	Nominal Wages (Rs.) (a)	Consumer Price Index (b)	Real Wages (Rs.) $(c = \frac{a}{b} \times 100)$	Real Wage Index (base 1985 = 100) $(d = \frac{c}{a, base} \times 100)$
1985	900	100	900.00	100.00
1986	950	108	879.63	97.74
1987	1020	117	871.79	96.87
1988	1050	131	801.53	89.06
1989	1100	148	743.24	82.58
1990	1200	155	774.19	86.02
1991	1275	175	728.57	80.95
1992	1400	180	777.78	86.42

When the actual time series i.e., nominal wages are deflated using the consumer price index, we find that real wages have been generally decreasing though the nominal wages are increasing. The situation is very well reflected by the real wage index.

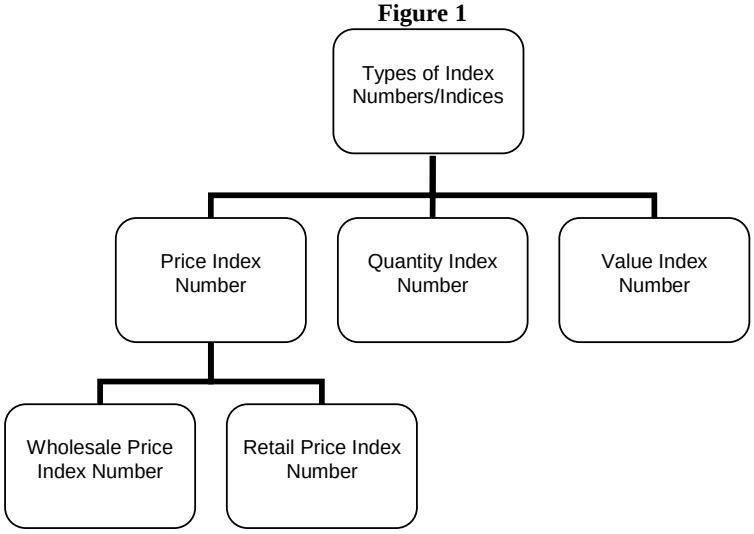
Limitations of Index Numbers

- The possibility of occurrence of errors at each stage of introduction of an index number makes it inefficient.
- Since index numbers are based on the sample data, they are only approximate indicators and may not exactly represent the changes in the relative level of a phenomenon.
- Since they are based on the sample data there is a possibility that they may not represent the exact change in the price level.
- The change in the customs, habits and tastes of people may make the index numbers not suitable for the present data.
- The averages those are used for the construction of index numbers have their own limitations. This is the reason for ineffective representation of index numbers.
- By selecting a suitable year as the base year or a suitable choice of commodities, price and quantity quotations, selfish and unscrupulous persons may get the desired result.
- There may be an error in each index number because there is no formula for measuring price change. Thus, there can be a formula error, which limits it to be an efficient representative.

TYPES OF INDEX NUMBERS

There are three types of principal indices. They are, the price index, the quantity index and the value index. These are represented in the following figure:

Figure 1



Price Index Number

The most frequently used form of index numbers is the price index. A price index compares changes in price from one period to another. Let us consider the price of edible oils. If an attempt is being made to compare the prices of edible oils this year to the prices of edible oils last year, it involves, firstly, a comparison of two price situations over time and secondly, the heterogeneity of the edible oils given the various varieties of oils. By constructing a price index number, we are summarizing the price movements of each type of oil in this group of edible oils into a single number called the price index. The Wholesale Price Index (WPI) and Retail Price Index are the different types of such indices. Consumer Price Index (CPI) , which is used for price indices is a popular measure of Retail Price Index.

Quantity Index Number

A quantity index measures the changes in quantity from one period to another. If in the above example, instead of the price of edible oils, we are interested in the quantum of production of edible oils in those years, then we are comparing quantities in two different years or over a period of time. It is the quantity index that needs to be constructed here. The popular quantity index used in this country and elsewhere is the Index of Industrial Production (IIP). The index of industrial production measures the increase or decrease in the level of industrial production in a given period compared to some base period.

Value Index Number

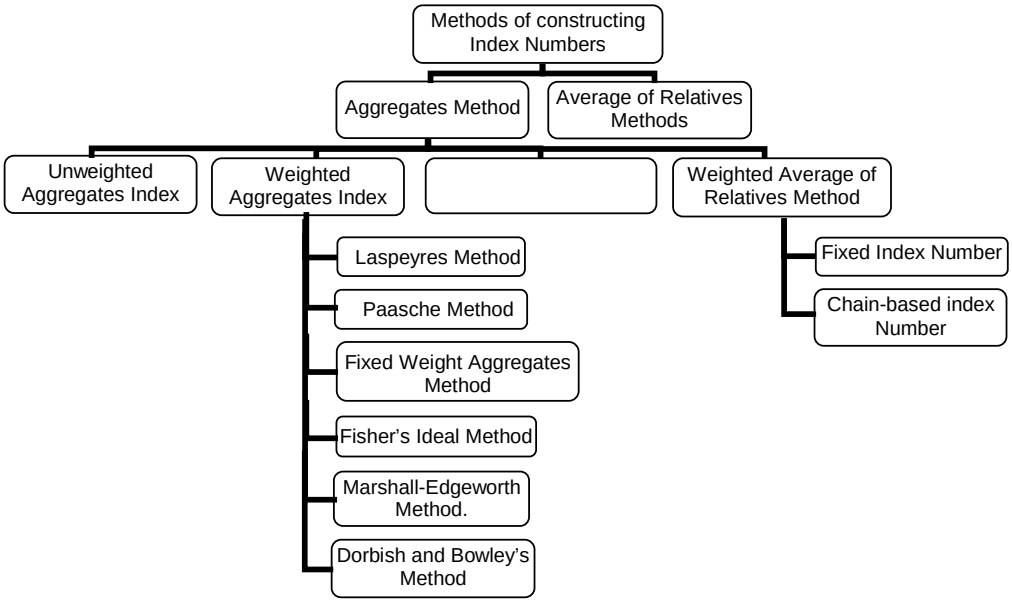
The value index is a combination index. It combines price and quantity changes to present a more spatial comparison. The value index as such measures changes in net monetary worth. Though the value index enables comparison of value of a commodity in a year to the value of that commodity in a base year, it has limited use. Usually, value index is used in sales, inventories, foreign trade, etc. Its limited use is owing to the inability of the value index to distinguish the effects of price and quantity separately.

METHODS OF CONSTRUCTING INDEX NUMBERS

There are two approaches for constructing an index number, namely the aggregates method and average of relatives method. The index constructed in either of these methods could be an unweighted index or a weighted index. A weighted index is an index where weights are assigned to the various items

constituting the index. On the other hand, an unweighted index is an index where equal weights are implicitly assigned.

Figure 2



1. Aggregates Method
   - Unweighted aggregates Index.
   - Weighted aggregates Index.
   - Laspeyres Method
   - Paasche's Method
   - Fixed Weight Aggregates Method.
   - Fisher's Ideal Method.
   - Marshall-Edgeworth Method.
   - Dorbish and Bowley's Method.
2. Average of Relatives Method
   - Unweighted Average of Relatives Method.
   - Weighted Average of Relatives Method.
   - Fixed Base Index Number.
   - Chain-Based Index Number.

Problems in the Construction of Index Numbers

- Selection of data is the first problem in the construction of index numbers. Utmost care should be taken while selecting the data for the purpose of construction of index numbers. For calculating index number, only standardized items are to be selected.
- The second problem that is posed in the construction of an index number is the selection of base period. Accurate base period gives authentic index number for comparison purposes. Hence, proper base must be selected in the construction of index numbers.
- The use of appropriate average is the next problem in the construction of index numbers. As stated earlier, index number is a special type of average and hence its calculation involves the use of appropriate averages. Thus, appropriate average helps to get authentic index number.

- Selection of weights has a considerable role to play in the construction of index numbers. Weights refer to the relative value of the different items in the construction of index numbers. All items are not of equal values; therefore, some weights are to be assigned. It is difficult to take correct decisions about correct weights.
- Selection of appropriate formula is of high significance in constructing the index number. Appropriate formula makes index number authentic and vice-versa.
- Index numbers are constructed with regard to price or quantity or any other measure. It is difficult to take the decision with regard to the variables to be measured.

AGGREGATES METHOD

Under the aggregates method of constructing an index number, we have unweighted aggregates index and the weighted aggregates index.

Unweighted Aggregates Index

An unweighted aggregates index is calculated by totalling the current year/given year's elements and then dividing the result by the sum of the same elements during the base period. To construct a price index, the following mathematical formula may be used

$$\text{Unweighted Aggregates Price Index } P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

Where,

$\sum P_1$ = Sum of all elements in the composite for current year.

$\sum P_0$ = Sum of all elements in the composite for base year.

This is the simplest method of constructing index numbers. The example demonstrates the application of an unweighted index.

Construction of Unweighted Price Index

Elements in the Composite	Prices (in Rs.)	
	2000 (P_0)	2001 (P_1)
Oranges (1 dozen)	20	28
Milk (1 liter)	5	8
LPG Cylinder	76	100
	101	136
Unweighted aggregates price index $= \frac{\sum P_1}{\sum P_0} \times 100$ $= \frac{136}{101} \times 100 = 134.65$		

Above, we measured changes in general price levels on the basis of changes in the prices of a few items. While the year 2000 was taken as the base year, a comparison has been made between the prices of 2001 and that of the base year 2000. As evident, the price index was 134.65 which means that the prices rose by 34.65 percent from 2000 to 2001. By no means should this price index be interpreted as a reflection of the price changes of all goods and services as this calculation is a rough estimate. On inclusion of other items/elements and varying weights in the composite, with 2000 as the base year and 2001 as the current year, there is every possibility that the calculated price index would be different from the price index calculated earlier. This factor can be cited as one of the drawbacks

of the simple unweighted index. The unweighted index does not reflect the reality since the price changes are not linked to any usage/consumption levels. On the other hand, a weighted index attaches weights according to their significance and hence is preferred to the unweighted index.

To make this clear, let us calculate the price index with the same data provided above, but by changing the milk consumption from 1 liter to 100 liters. The following table provides the calculation of the price index:

Calculation of Unweighted Price Index

Elements in the Composite	Prices (in Rs.)	
	2000 (P_0)	2001 (P_1)
Oranges (1 dozen)	20	28
Milk (100 liters)	500	800
LPG Cylinder	76	100
	596	928

Unweighted aggregates price index $= \frac{\sum P_1}{\sum P_0} \times 100$

$= \frac{928}{596} \times 100 = 155.70$

Merely by changing the milk consumption in the composite, the unweighted price index changed from 134.65 to 155.70. As a result of ensuring that equal importance is given to all items in the composite irrespective of the consumption, the unweighted aggregates never gained much acceptance.

An unweighted aggregates quantity index and, an unweighted aggregates value index can be calculated on similar lines as calculated for price index. A mere substitution of quantities or values for prices in the equation $(\sum P_1 / \sum P_0) \times 100$ would suffice.

Weighted Aggregates Index

In a weighted aggregates index, weights are assigned according to their significance. Consequently, the weighted index improves the accuracy of the general price level estimate based on the calculated index. Generally, the level of consumption is taken as a measure of its importance in computing a weighted aggregates index. There are various methods of assigning weights to an index. The more important ones are:

- Laspeyres Method.
- Paasche's Method.
- Fixed Weight Aggregates Method.
- Fisher's Ideal Method.
- Marshall-Edgeworth Method.
- Dorbish and Bowley's Method.

LASPEYRES METHOD

Laspeyres method uses the quantities consumed during the base period in computing the index number. This method is also the most commonly used

method, which incidentally requires quantity measures for only one period. Laspeyres index can be calculated using the following formula:

$$\text{Laspeyres Price Index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Where,

p_1 = Prices in the current year.

p_0 = Prices in the base year.

q_0 = Quantities in the base year.

Laspeyres Index

Commodities	Production (in quintals)		Price per quintal (in Rs.)		$p_0 \times q_0$	$p_1 \times q_0$	$p_0 \times q_1$
	q_0	q_1	p_0	p_1			
	1995	2000	1995	2000			
Rice	46.60	58.00	700	910	32,620.00	42,406.00	40,600.00
Sugar	14.57	17.92	620	950	9,033.40	13,841.50	11,110.40
Jowar	69.46	85.10	205	300	14,239.30	20,838.00	17,445.50
Wheat	33.84	40.30	330	470	11,167.20	15,904.80	13,299.00
					67,059.90	92,990.30	82,454.90

$$\text{Laspeyres price index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{92,990.30}{67,059.90} \times 100$$

$$= 138.67$$

In general, Laspeyres price index calculates the changes in the aggregate value of the base year's list of goods when valued at current year prices. In other words, Laspeyres index measures the difference between the theoretical cost in a given year and the actual cost in the base year of maintaining a standard of living as in the base year.

Also, Laspeyres quantity index can be calculated by using the formula,

$$\text{Laspeyres Quantity Index} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

Where,

q_1 = Quantities in the current year and q_0, p_0 are as defined earlier.

Using the same data as provided in the above table, Laspeyres quantity index is

$$= \frac{82,454.90}{67,059.90} \times 100 = 122.96$$

A Laspeyres index is simpler in calculation and can be computed once the current year prices known as the weights are base year quantities in a price index. This also enables easy comparability of one index with another. Interestingly, Laspeyres tends to overestimate the rise in prices or has an upward bias.

- There is usually a decrease in the consumption of those items for which there has been a considerable price hike and the usage of base year quantities will result in assigning too much weight to prices that have increased the most and the net result is that the numerator of the Laspeyres index will be too large.
- Similarly, when the prices go down, consumers tend to demand more of those items that have declined the most and hence the usage of base period quantities will result in too low weight to prices that have decreased the most and the net result is that the numerator of the Laspeyres index will again be too large.

This is a major disadvantage of the Laspeyres index.

However, the Laspeyres index remains most popular for reasons of its practicability. In most countries, index numbers are constructed by using Laspeyres formula.

PAASCHE'S METHOD

Paasche's index is similar to that of computing a Laspeyres index. The difference is that the Paasche's method uses quantity measures for the current period rather than for the base period. The Paasche's index can be calculated using the following formula:

$$\text{Paasche's Price Index} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Where,

p_1 = Prices in the current year.

p_0 = Prices in the base year.

q_1 = Quantities in the current year.

The table below represents the calculation of Paasche's index. In general, Paasche's index reflects the change in the aggregate value of the current year's (given period's) list of goods when valued at base period prices. Paasche's index is not frequently used in practice when the number of commodities is large. This is because for Paasche's index, revised weights or quantities must be computed for each year examined. Such information is either unavailable or hard to gather adding to the data collection expense which makes the index unpopular.

Paasche's Index

	2000		2001					
Commodities	Price	Quantity	Price	Quantity	$p_0 q_0$	$p_0 q_1$	$p_1 q_0$	$p_1 q_1$
	(p_0)	(q_0)	(p_1)	(q_1)				
A	3	18	4	15	54	45	72	60
B	5	6	5	9	30	45	30	45
C	4	20	6	26	80	104	120	156
D	1	14	3	15	14	15	42	45
					178	209	264	306

Paasche's Price Index $= \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = 146.41$

Paasche's Quantity Index $= \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100 = 115.91$

Paasche's price index is calculated as 146.41. Let us calculate the price index by the Laspeyres method using the same data.

$$\text{Laspeyres Price Index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{264}{178} \times 100$$

$$= 148.31$$

The difference between the Paasche's index and Laspeyres index reflects the change in consumption patterns of the commodities A, B, C and D used in that table. As the weighted aggregates price index for the set of prices was 148.31 using the Laspeyres method and 146.41 using the Paasche's method for the same set, it indicates a trend towards less expensive goods.

Generally, Laspeyres and Paasche's methods tend to produce opposite extremes in index values computed from the same data. The use of Paasche's index requires the continuous use of new quantity weights for each period considered. As opposed to the Laspeyres index, Paasche's index generally tends to underestimate the prices or has a downward bias.

Since people tend to spend less on goods when their prices are rising, the use of the Paasche's which bases on current weighting, produces an index which does not estimate the rise in prices rightly showing a downward bias. Since all prices or all quantities do not move in the same order, the goods which have risen in price more than others at a time when prices in general are rising, will tend to have current quantities relatively smaller than the corresponding base quantities and they will thus have less weight in the Paasche's index.

FIXED WEIGHT AGGREGATES METHOD

In fixed weight aggregates method, the weights used are neither from base period nor from current period but from a representative period. These weights are generally referred to as representative weights or as fixed weights. These fixed weights are unaffected by the selection of the base period. This is the advantage under this method. The user of the method will be able to select a base year that is convenient to him enabling him to change the price base yet retaining the fixed weights.

The students may refer to the weights assigned to various industry groups constituting the Index of Industrial Production presented in the annexure.

FISHER'S IDEAL METHOD

Prof. Irving Fisher has proposed a formula for constructing index numbers, which is called the 'Fisher's Ideal Index'. The Ideal index is given by the following formula:

$$\text{Fisher's Ideal Index} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

As evident from the above formula,

Fisher's Ideal Index is the geometric mean of the Laspeyres and Paasche's indices.

The following advantages can be said to be in favor of Fisher's Ideal Index:

1. Theoretically, geometric mean is considered the best average for the construction of index numbers and Fisher's index uses geometric mean.
2. As already noted, Laspeyres index and Paasche's index indicate opposing characteristics and Fisher's index reduces their respective biases. In fact, Fisher's ideal index is free from any bias. This has been amply demonstrated by the time reversal and factor reversal tests.
3. Both the current year and base year prices and quantities are taken into account by this index.

Fisher's Ideal Index

Commodities	2000		2001		$p_0 q_0$	$p_0 q_1$	$p_0 q_1$	$p_1 q_1$
	Price/kg	Qty	Price/kg	Qty				
	(p_0)	(q_0)	(p_1)	(q_1)				
Tea, Grade I	78	7	85	10	546	595	780	850
Tea, Grade II	69	5	80	5	345	400	345	400
Tea, Grade III	62	4	72	6	248	288	372	432
					1139	1283	1497	1682

$$\text{Fisher's Ideal Index} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{1283}{1139} \times \frac{1682}{1497}} \times 100 = 112.50$$

The index is not widely used owing to the practical limitations of collecting data. Fisher's Ideal Quantity Index can be found out by the formula,

$$\sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} x \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100$$

MARSHALL-EDGEWORTH METHOD

Marshall-Edgeworth method uses both the current year as well as the base year prices and quantities. Marshall-Edgeworth Index can be computed using the following formula:

$$\text{Marshall-Edgeworth Index} = \frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0} \times 100$$

Where,

p_0, q_1, p_0 and p_1 follow the usual notations.

Marshall-Edgeworth Index

Commodities	Base Year		Current Year				
	Price (p_0)	Qty (q_0)	Price (p_1)	Qty (q_1)	$[(q_0 + q_1)]$	$[p_0(q_0 + q_1)]$	$[p_1(q_0 + q_1)]$
A	7	17	13	25	42	294	546
B	6	23	7	25	48	288	336
C	11	14	13	15	29	319	377
D	4	10	8	8	18	72	144
						973	1403
Marshall-Edgeworth Index $= \frac{1,403}{973} x 100 =$ $= 144.19$							

Marshall-Edgeworth Index is simple to construct, but suffers from the problems in data collection. However, the Marshall-Edgeworth index closely approximates the results obtained by the Fisher's Ideal index.

DORBISH AND BOWLEY'S METHOD

Dorbish and Bowley suggest simple arithmetic mean of Laspeyres and Paasche's indices mentioned above. The arithmetic mean of two indices was suggested for taking into account the influence of both the current and the base period. The formula is given below:

$$P_{01} = \frac{L + P}{2}$$

Where,

L – Laspeyres Index, and

P – Paasche's Index

$$\text{or } p_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100$$

AVERAGE OF RELATIVES METHOD

We have seen the construction of an index number using the aggregates method. In this section, we shall see the construction of an index using the average of relatives method.

Unweighted Average of Relatives Method

Let us begin with a price index. When a price index is constructed, all price relatives are to be obtained for all the items included in the index after which the average of price relatives is obtained using any one of the measures of central tendency namely, arithmetic mean, geometric mean, median, mode or harmonic mean. A price relative may be generally understood as the ratio of the price of a single item in a given period to its price in the base period.

Though theoretically, any measure of central tendency can be used to obtain the index, the general practice is to use arithmetic mean for averaging the price relatives. The price index using the average of relatives method can be constructed using the following formula:

$$\text{Unweighted average of relatives index} = \frac{\sum\left(\frac{p_1}{p_0} x 100\right)}{n}$$

Where,

P_1 = Prices in the current/given year

P_0 = Prices in the base year

n = Number of products/items in the composite

The ratio P_1/P_0 is the price relative.

Unweighted Average of Relatives Index

Elements in the Composite	Prices/Kg		$\frac{p_1}{p_0} \times 100$
	Jan, 20 x 0 (p_0)	June, 20 x 0 (p_1)	
Rice	8.50	9.50	111.76
Wheat	4.75	5.00	105.26
Salt	3.00	3.00	100.00
Sugar	9.00	11.00	122.22
			439.24

$$\text{Unweighted average of relatives price index} = \frac{\sum\left(\frac{p_1}{p_0} X 100\right)}{n} = \frac{439.24}{4}$$

$$= 109.81$$

On similar lines, quantity index using average of relatives method can be computed with the help of the following formula:

$$\text{Unweighted Average of Relatives Quantity Index} = \frac{\sum\left(\frac{q_1}{q_0} x 100\right)}{n}$$

Where,

q_1 = Quantities in the given period.

q_0 = Quantities in the base year

n = Number of elements in the composite

The price index or the quantity index computed by the average of relatives method would be the same regardless of the way in which the prices are quoted or quantities are measured. In other words, price/quantity relatives are pure numbers and hence free from the units of measurement.

Also, the average of relatives method converts each element in the composite to a relative scale where each element is expressed in percentages and measured against a base of 100. The only impediment to such an index being constructed is the selection of an appropriate average. In general, arithmetic mean is used to take the average of the price relatives. As such, index is not influenced by extreme items. But the use of arithmetic mean, though simple and easy to understand, has a major disadvantage in that there is a tendency to overemphasize increases and undervalue decreases in prices. Though the use of geometric mean would overcome these tendencies, it is difficult to compute and its usage is avoided for this reason. The unweighted average of relatives method suffers from one or more limitations. The relatives (price/quantity) are assumed to have equal importance. As some relatives are economically more significant than others, assigning of equal weightages is undesirable.

Weighted Average of Relatives Method

The weighted average of relatives method is most commonly used than the unweighted average of relatives method. The value of each element in the composite is used as the weight in this method. Values could be from the base year, current/given year or any fixed period.

The computation of a weighted average of relatives index is done with the help of the following formula:

$$\text{Weighted Average of Relatives Price Index} = \frac{\sum\left[\left(\frac{p_1}{p_0} x 100\right)(p_n q_n)\right]}{\sum p_n q_n}$$

Where,

p_1 = Prices in the current year

p_0 = Prices in the base year

q_n = Prices in the fixed period (could also be base or current period).

q_n = Quantities in the fixed period (could also be base or current period).

$p_n q_n$ = Value in the fixed period (can be replaced by $p_0 q_0$ or $p_1 q_1$ as the case may be).

If the weighted average of relatives price index is to be calculated with the base values, the formula to be used is as follows:

Weighted average of relatives price index

$$= \frac{\sum\left[\left(\frac{p_1}{p_0} x 100\right)(p_0 q_0)\right]}{\sum p_0 q_0}$$

$$= \frac{\sum \frac{p_1}{p_0} x\, p_0 q_0 \, x 100}{\sum p_0 q_0}$$

$$= \frac{\sum p_1 q_0}{\sum p_0 q_0} x 100,$$

Which incidentally is the weighted aggregate price index using Laspeyres method.

The weighted average of price relatives using base values is the same as the weighted aggregate of actual prices (Laspeyres method) only when the arithmetic mean is used for averaging the relatives and the base year values are used as weights. If either of these conditions are not satisfied, such a comparison would not be possible.

Weighted Average of Relatives Price Index

Elements in the Composite	Production (lakh tonnes)		Weights	Price Relative	Base Value	Weighted Percentage Relative
	1952-53	1965-66	1952-53			
	p_0	p_1	q_0	$\frac{p_1}{p_0} \times 100$	$p_0 q_0$	
(1)	(2)	(3)	(4)	(5)	(6)	(7) = (5) x (6)
Rice	234.2	306.1	33.98	130.70	7958.12	1040126.3
Wheat	67.6	107.2	12.16	158.58	822.02	130355.93
Jowar	60.4	75.0	4.86	124.17	293.54	36448.86
Gram	38.0	44.0	3.58	115.79	136.04	15752.07
Maize	29.2	36.0	3.06	123.29	89.35	11015.96
Sugarcane	52.6	118.3	7.01	224.90	368.73	82927.38
Cotton	16.5	14.3	3.01	86.67	49.67	4304.90
					9717.47	1320931.40

Weighted Average of Relatives Price Index $= \frac{\sum(\frac{p_1}{p_0} x 100)(p_0 q_0)}{\sum p_0 q_0}$

$= \frac{1320931.40}{9717.47} = 135.93$

Generally, either base values or fixed values are used when computing a weighted average of relatives index. This is so because, when current values are used in computing a weighted average of relatives price index, comparison of values from different time periods is not possible owing to the change in the prices and the quantities in these time periods.

The advantage of a weighted average of relatives index is that the price or quantity relatives for each element in the aggregate are themselves a simple index that can be used for analysis. Also, different indices constructed with the same base using the average of relatives method can be combined to form a new index.

CHAIN INDEX NUMBERS

So far, we have constructed index numbers with a fixed base. Sometimes, comparison between the current/given year and the base year becomes meaningless once time elapses making the base year remote. An index can be made more representative by using the chain base method of constructing index numbers. In this method, the base year is not fixed but changes from year to year. Here, new relatives are constructed for each year with the previous year as base. The new relatives enable comparison between one period and another period, which succeeds it and hence referred to as 'Link relatives'. When these link relatives are associated to a common base, and chained together by successive multiplication, it results in the formation of a chain index number.

Chain index for a given year

$$= \frac{\text{Average Link Relative of the Given Year x Chain Index of Previous Year}}{100}$$

Where,

$$\text{Link Relative} = \frac{\text{Price in a given period}}{\text{The previous year's price}} \times 100$$

The construction of a chain index has been illustrated in the following example.

Illustration 1

Construct chain index numbers for the years 1998-99 to 2001-2002 from the following data:

Index Numbers of Wholesale Prices

			1981-82 = 100	
	Primary Articles	Fuel Group	Manufactured Products	All Commodities
	(A)	(B)	(C)	(D)
1998-1999	163.6	156.6	168.6	165.7
1999-2000	184.9	175.8	182.8	182.7
2000-2001	218.4	199.0	203.5	207.8
2001-2002	234.6	227.1	225.6	228.7

Solution

Construction of Chain Indices

	1998-99	1999-2000	2000-2001	2001-2002
A	100	$\frac{184.9}{163.6} \times 100$ = 113.02	$\frac{218.4}{184.9} \times 100$ = 118.12	$\frac{234.6}{218.4} \times 100$ = 107.42
B	100	$\frac{175.8}{156.6} \times 100$ = 112.26	$\frac{199.0}{175.8} \times 100$ = 113.20	$\frac{227.1}{199.0} \times 100$ = 114.12
C	100	$\frac{182.8}{168.6} \times 100$ = 108.42	$\frac{203.5}{182.8} \times 100$ = 111.32	$\frac{225.6}{203.5} \times 100$ = 110.86
D	100	$\frac{182.7}{165.7} \times 100$ = 110.26	$\frac{207.8}{182.7} \times 100$ = 113.74	$\frac{228.7}{207.8} \times 100$ = 110.06
Total of Link relatives	400	443.96	456.38	442.46
Average of Link relatives	100	110.99	114.10	110.62
Chain index	100	$\frac{110.99 \times 100}{100}$ = 110.99	$\frac{114.10 \times 110.99}{100}$ = 126.64	$\frac{110.62 \times 126.64}{100}$ = 140.09

The following steps are to be noted in the construction of a chain index:

i. **Obtain link relatives**. These are to be expressed for each year as percentages of the preceding year.

In the above example, link relative for the year 1999-2000 for (Primary articles) A is

$$= \frac{\text{Price of A in 1999-2000}}{\text{Price of A in 1998-99}} \times 100$$

$$= \frac{184.9}{163.6} \times 100 = 113.02$$

ii. **Obtain the Average Link Relative of the Year**. This is the ratio of the total link relatives for each year to the number of items in the index. In the above example, average link relative for the year 1999-2000, is Total of link relatives for 1999-2000/4 (443.96/4) = 110.99

iii. **Successive Multiplication of Average Link Relatives to be Done**. For any year, the chain index is the product of the average link relative of that year and the chain index of the previous year, divided by 100.

In the above example, chain index for the year 2001-2002 is

$$\frac{\text{Average of Link Relative for 1992-93} \times \text{Chain Index for 2000-2001}}{100}$$

$$= \frac{110.62 \times 126.64}{100} = 140.09.$$

Advantages of Chain Base Method

- The link relatives calculated by using the chain base method, enable comparisons over successive years. There is not much significance in business to the comparisons with the current period and remote past (the base year) and as such the usage of chain index numbers is ideal.
- Chain base method enables the introduction of new items and the deletion of old items without altering the original series. It is thus flexible.
- Whenever found necessary, weights can be adjusted in chain base method.
- Seasonal variations have minimal impact on chain index numbers.

Disadvantage of Chain Base Method

As percentages are chained together in chain index numbers, long range comparisons are not very accurate.

VALUE INDEX NUMBERS

The value index number as described earlier is a combination index which combines price and quantity changes. In view of the difficulties experienced in price and quantity indices, the usage of the value index has been suggested. The value index number can be calculated by the following formula:

$$= \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$$

where, p_1 is the price level of the current year, q_1 is the quantity level of the current year, p_0 is the price level of the base year and q_0 is the quantity level of the base year.

Interestingly, in value indices, weights need not be used as they are inherent in these indices. The value index, an aggregate of all values, measures the changes in values in the base year and the values in the given year. The value index can also

be obtained by the product of price and quantity indices. Let us calculate the value of some equity shares owned by an investor on July 27, 20x1 relative to the value of the shares owned by him on Jan 8, 20x0 as measured by the value index.

Value Index

	As on Jan 8, 20x0			As on July 27, 20x1		
Equity Shares	Price in Rs. (p_0)	No. of Shares (q_0)	Value (p_0q_0)	Price in Rs. (p_1)	No. of Shares (q_1)	Value (p_1q_1)
Premier Auto	46.25	300	13,875	37.00	100	3,700
Mysore Breweries	230.00	200	46,000	362.50	100	36,250
Fuller KCP	240.00	200	48,000	395.00	200	79,000
Goodlass Nerolac	145.00	100	14,500	160.00	200	32,000
			1,22,375			1,50,950

The value index = $\frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100 = \frac{1,50,950}{1,22,375} \times 100 = 123.35$

The value of the equity shares owned by the investor has gone up by 23.35 percent.

TESTS FOR CONSISTENCY

The consistency of the index numbers has been tested over the years. The most important of these tests are:

a. The time reversal test.

b. The factor reversal test.

c. The circular test.

Time Reversal Test

Time reversal test was developed by Prof. Irving Fisher. The test implies that for a price/quantity index, if the time periods are reversed, the resulting index should be the reciprocal of the original price/quantity index.

If $I_{0,1}$ denotes the original index for the current year with a given base year, and if $I_{1,0}$ denotes the resulting index, with time periods reversed, for the base year with the current year as a base year, then according to the time reversal test,

$$I_{0,1} = \frac{1}{I_{1,0}} \text{ (or) } I_{0,1} \times I_{1,0} = 1$$

Where,

$$I_{0,1} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}},$$

$$I_{1,0} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} \text{ and}$$

$p_1 q_1$, $p_0 q_0$ refer to usual notations.,

It is noteworthy in this context that Fisher's Ideal Index, Marshall-Edgeworth Index and Fixed Weights Aggregate Index satisfy this test. Laspeyres Index and Paasche Index do not satisfy it.

Factor Reversal Test

This test was also suggested by Prof. Irving Fisher. According to Prof. Fisher, just as each formula should permit the interchange of the two time periods without giving inconsistent results, it ought to permit interchanging of the prices and quantities without giving inconsistent result, i.e., the two results multiplied together should give the true value ratio.

The product of change in prices in the current year and the change in quantities in the current year should be equal to $V_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$

Where,

p_1 = Prices in the current year.

p_0 = Prices in the base year.

q_1 = Quantities in the current year.

q_0 = Quantities in the base year.

or

$$P_{01} \times Q_{01} = V_{01}$$

or

$$\frac{P_{01} \times Q_{01}}{V_{01}} = 1$$

Where,

P_{01} = Price index for the given year with reference to the base year = $P_{01} = \sqrt{\frac{p_1 q_0}{p_0 q_0} \times \frac{p_1 q_1}{p_0 q_1}}$

Q_{01}= Quantity index for the given year with reference to the base year = $q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$

V_{01} = Value Factor

The test can be applied to the index numbers by interchanging p to q and q to p.

Except Fischer's Ideal Index, all other elementary indices, simple as well as weighted, fail to satisfy this test.

Illustration 2

The following figures relate to the prices and quantities of certain commodities:

Commodity	2005		2006	
	Price (Rs.)	Quantity (in Kgs.)	Price (Rs.)	Quantity (in Kgs.)
A	10	400	12	450
B	11	500	11	520
C	14	300	17	300
D	8	280	10	290
E	12	150	13	200

You are required to construct Fisher Ideal Index and show that it satisfies time reversal and factor reversal tests.

Solution

Construction of Fisher Ideal Index

Commodity	2005		2006		$\sum p_1q_0$	$\sum p_0q_0$	$\sum p_1q_1$	$\sum p_0q_1$
	Price (Rs.) p_0	Quantity (Kgs) q_0	Price (Rs.) p_1	Quantity (Kgs) q_1				
A	10	400	12	450	4,800	4,000	5,400	4,500
B	11	500	11	520	5,500	5,500	5,720	5,720
C	14	300	17	300	5,100	4,200	5,100	4,200
D	8	280	10	290	2,800	2,240	2,900	2,320
E	12	150	13	200	1,950	1,800	2,600	2,400
					$\sum p_1q_0$ = 20,150	$\sum p_0q_0$ = 17,740	$\sum p_1q_1$ = 21,720	$\sum p_0q_1$ = 19,140

$$\text{Fisher Ideal Index} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100$$

$$= \sqrt{\frac{20,150}{17,740} \times \frac{21,720}{19,140}} \times 100$$

$$= 1.135 \times 100 = 113.5$$

Time Reversal Test

Time reversal test is satisfied when $p_{01} \times p_{10}=1$

$$p_{10} = \sqrt{\frac{\sum p_0q_1}{\sum p_1q_1} \times \frac{\sum p_0q_0}{\sum p_1q_0}} = \sqrt{\frac{19,140}{21,720} \times \frac{17,740}{20,150}}$$

$$p_{01} \times p_{10} = \sqrt{\frac{20,150}{17,740} \times \frac{21,720}{19,140} \times \frac{19,140}{21,720} \times \frac{17,740}{20,150}} = 1$$

Hence time reversal test is satisfied.

Factor Reversal Test

Factor reversal test is satisfied when $p_{01} \times q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$

$$q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} = \sqrt{\frac{19,140}{17,740} \times \frac{21,720}{20,150}}$$

$$P_{01} \times q_{01} = \sqrt{\frac{20,150}{17,740} \times \frac{21,720}{19,140} \times \frac{19,140}{17,740} \times \frac{21,720}{20,150}} = \frac{21,720}{17,740}$$

$$\frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{21,720}{17,740}$$

Hence, factor reversal test is satisfied.

Circular Test

Circular test is an extension of the time reversal test. This test is used while measuring price changes over a number of years with the shifting of base occurring frequently. According to this test, an index constructed for the year 'X' on base year 'Y' and for the year 'Y' on base year 'Z' should yield the same result as an index constructed for 'X' on base year 'Z'.

i.e. $I_{0,1} \times I_{1,2} \times I_{2,0} = 1$.

Laspeyres, Paasche and Fisher's indices fail to satisfy this test. Only the fixed weight aggregates method and simple aggregates method satisfy this test.

The indices satisfying the circular test would be amenable to change from year to year without referring to the base year. The indices have the advantage of reduced computational work in the event of a change in the base year.

Note: This has been discussed theoretically only.

CONSUMER PRICE INDEX NUMBERS

The consumer price index number is also known as cost of living numbers. It represents the average change in prices paid by the consumer for specified goods or services. The general or wholesale price index number does not give an exact idea of the effect of change in general price level on the cost of living of different people. Preparing the consumer price index minimizes this shortcoming. Consumer price index number or Cost of living index number are designed to measure the effect of change in the general level of price of a basket of goods and services on the purchasing power of consumer during a particular point of time compared to base period. Due to variations in the tastes, customs and fashions, different classes of people consume different type of commodities and even the same type of commodities is consumed in different or same proportion. Accordingly, any change in price affect these consumers differently. Hence, consumer price index is prepared to determine and study the effect of rise or fall in the prices of various commodities consumed by particular group of people on their cost of living. The construction of such index is of great significance because the wages and salaries are adjusted in accordance with the consumer price index. Consumer Index Numbers are constructed for different classes of people and also for different geographical areas separately.

Utility of Consumer Price Indices

- They are used commonly in wage negotiations and wage contracts.
- Government uses index numbers for wage policy, price policy, taxation, rent control policy and for other economic policies.
- It also measures the changing purchasing power of currency.
- Market for particular goods and services are analyzed with the help of index number.

Table 1: Group/Sub-group wise percentage change in All-India index from last month and last year

						Base: 1984-85=100
	Feb-07	Jan-08	Feb-08	Last month	Last year	Group & Sub-group
I	487	513	517	0.78	6.16	I FOOD, BEVERAGES & TOBACCO
	480	505	508	0.59	5.83	a. Cereals
	611	597	595	–0.34	–2.62	b. Pulses
	337	385	400	3.90	18.69	c. Oils & Fats
	575	616	619	0.49	7.65	d. Meat, Fish etc.
	420	461	464	0.65	10.48	e. Milk & Milk Products
	598	626	619	–1.12	3.51	f. Condiment, Spices etc.
	480	474	480	1.27	0.00	g. Vegetables
	445	409	418	2.20	–6.07	h. Fruits
	336	309	312	0.97	–7.14	i. Sugar, Honey etc.
	510	536	540	0.75	5.88	j. Non-alc Beverages
	618	653	657	0.61	6.31	k. Prep. Meals etc.
	707	826	829	0.36	17.26	l. Pan, Supari, Tobacco tc.
II	563	590	592	0.34	5.15	II FUEL & LIGHT
III	529	550	550	0.00	3.97	III HOUSING
IV	436	453	455	0.44	4.36	IV CLOTHING, BEDDING & FOOT-WEAR etc.
	428	443	445	0.45	3.97	a. Clothing & Bedding
	496	525	527	0.38	6.25	b. Foot-wear
V	498	517	519	0.39	4.22	V MISCELLANEOUS
	511	541	543	0.37	6.26	a. Medical care
	392	410	410	0.00	4.59	b. Education
	418	436	436	0.00	4.31	c. Recreation&Amusement
	687	696	700	0.57	1.89	d. Transport&Communicaton
	390	412	415	0.73	6.41	e. Personal Care & Effect
	433	453	457	0.88	5.54	f. Household Requisites
	587	613	616	0.49	4.94	g. Others
	497	520	523	0.58	5.23	General Index

Source: http://mospi.nic.in/mospi_cpi.htm

Main Steps in Construction of Consumer Price Index Numbers

- **Scope and Coverage:** The first step in the construction of cost of living is to identify the class of people for whom the index is to be constructed. It is absolutely essential to decide the class of people such as industrial worker, agricultural worker, labor class etc., for whom the index is desired. In addition to the class of people, it is also necessary to decide the geographical area such as urban area, rural area, city, town etc., to be covered by the index.
- **Family Budget Enquiry:** Once the scope and coverage of the index is clearly defined, the next step is to conduct a sample of family budget enquiry covering the population group for whom the index is designed. The object of the enquiry is to determine the amount/expenses, which an average family included in the group spends/incurs on different items of consumption. Such an enquiry is to be conducted in the normal period of economic stability and on random basis. The following information is provided by enquiry: (1) Nature, quality and quantity of the commodities consumed by a particular class of people. These commodities are broadly classified as – Food, Clothing, Fuel and Lighting, House rent and Miscellaneous. Each of these groups are further sub-divided into smaller groups known as sub-group. (2) Obtaining the price quotations i.e., the retail prices of different commodities for whom the index is designed. Such price quotations for the selected commodities are to be collected from the localities in which the particular class of people resides or from the shops, societies from which they usually do their purchases. (3) From the commodities prices and quantity consumer, we can obtain – the expenditure on each item as a % of total expenditure of the whole group and also the expenditure on each group as a % of the total expenditure of all the groups.
- **Methods for Construction of Cost of Living Index:** For different class of people, the relative importance of different items of consumption is different. It is even different within the same class from one region to another. Depending on the relative importance of different consumption items, the cost of living index is taken as weighted indices. Following methods are applied for the construction of consumer price index.
- **Aggregate Expenditure Method:** It is also known as Weighted Aggregate Method. In this method, the quantities of commodities consumed by a particular group in the base year are used as weights. The prices of various commodities for the current year are multiplied by the quantity consumed in the previous/base year and the aggregate expenditure is obtained. Similarly, the prices of the base year are multiplied by the base year's quantities and aggregate expenditure for the base year is obtained. The consumer price index is obtained by dividing the aggregate expenditure of the current year by the aggregate expenditure of the base year multiplied by 100. Symbolically,

$$\text{Consumer Price Index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

  This method is similar to Laspeyres method and is popularly used for constructing consumer price index.
- **Family Budget Method:** This method is also known as Method of Weighted Relatives. In this method, the aggregate expenditure of an average family or the value of quantities consumed constitutes weight. Weights are obtained by multiplying the quantity consumed with its price. Then the price relatives are obtained, which are multiplied by the value weights for each item. The product so obtained is divided by the sum of the weights. Symbolically,

$$\text{Consumer price index} = \frac{\sum PV}{\sum V} \text{ or } \frac{\sum WI}{\sum W}$$

$$= \text{ where V or W } = p_0 q_0 \quad \frac{\sum WI}{\sum W}$$

$$\text{P or I} = \frac{P_1}{P_0}$$

Illustration 3

Construct the Cost of Living Index Number from the Table Given Below:

Group	Index for 2003	Expenditure
Food	550	46%
Clothing	215	10%
Fuel & Lighting	220	7%
House rent	150	12%
Miscellaneous	275	25%

Solution

Construction of Cost of Living Index Number

Group	Index (I)	Expenditure (W)	WI
Food	550	46	25300
Clothing	215	10	2150
Fuel & Lighting	220	7	1540
House rent	150	12	1800
Miscellaneous	275	25	6875
		$\sum W = 100$	$\sum WI = 37665$

$$\text{Cost of Living Index} = \frac{\sum WI}{\sum W} = \frac{37665}{100} = 376.65$$

Illustration 4

An enquiry into the budget of middle class families in Mumbai gave the following information:

Group	Price 2002	Price 2003	Expenditure
Food	150	174	35%
Clothing	100	125	20%
Fuel	20	25	10%
House rent	50	60	15%
Miscellaneous	60	90	20%

Construct the cost of living index for an average family in Mumbai.

Solution

Construction of Cost of Living Index

Group	p_0	p_1	$\frac{p_1}{p_0}\times 100$	W	PW
Food	150	174	116	35	4060
Clothing	100	125	125	20	2500
Fuel	20	25	125	10	1250
House rent	50	60	120	15	1800
Miscellaneous	60	90	150	20	3000
				$\sum W=100$	$\sum PW$ =12610

$$\text{Cost of Living Index} = \frac{\sum PW}{\sum W} = \frac{12610}{100} = 126.10.$$

ADDITIONAL ILLUSTRATIONS

Illustration 1

The sales manager of Alpha Company is examining the commission rate employed for the last 3 years. Below are the commission earnings of the company's top five sales persons. (Commission earnings are in rupees)

	1993	1994	1995
A	48,500	55,100	63,800
B	41,900	46,200	60,150
C	38,750	43,500	46,700
D	36,300	45,400	39,900
E	33,850	38,300	50,200

Using 1993 as the base period, express the commissions earnings in 1994 and 1995 in terms of the unweighted aggregates index.

Solution

Calculation of Unweighted Aggregate Index

	1993	1994	1995
	p_0	p_1	p_2
A	48,500	55,100	63,800
B	41,900	46,200	60,150
C	38,750	43,500	46,700
D	36,300	45,400	39,900
E	33,850	38,300	50,200
Total	1,99,300	2,28,500	2,60,750

Unweighted Aggregate Index = $\frac{\sum P_1}{\sum P_0} \times 100$

Index for 1993 = $\frac{1,99,300}{1,99,300} \times 100 = 100$

Index for1994 = $\frac{2,28,500}{1,99,300} \times 100 = 114.6513$

Index for 1995 = $\frac{2,60,750}{1,99,300} \times 100 = 130.8329$

Illustration 2

Telly Vision is a player in the television industry. The selling price and the sales figures for the past 3 years are as follows:

	Selling price (in Rs)			Units sold		
Model	1995	1996	1997	1995	1996	1997
Vision	17,000	17,900	18,600	11,200	12,000	12,000
Emperor	13,100	13,600	13,700	8,000	9,600	11,200
Prince	11,200	11,400	11,900	9,600	10,400	12,000
Crown	8,500	8,700	8,800	11,200	10,400	12,000
Pearl	5,500	5,600	5,800	12,000	14,400	18,000

a. Using 1995 as the base year calculate the Laspeyres, Paasche's and Fisher's price indices for the year 1996 and 1997.

b. Calculate the unweighted average of relatives price index for the year 1996 and 1997 with 1995 as the base year.

Solution

Calculation of Laspeyres Index

Model	1995 p_0	1996 p_1	1997 p_2	1995 q_0	p_0q_0	p_1q_0	p_2q_0
Vision	17	17.9	18.6	11.2	190.4	200.48	208.32
Emperor	13.1	13.6	13.7	8.0	104.8	108.8	109.60
Prince	11.2	11.4	11.9	9.6	107.52	109.44	114.24
Crown	8.5	8.7	8.8	11.2	95.2	97.44	98.56
Pearl	5.5	5.6	5.8	12.0	66.0	67.2	69.6
					563.92	583.36	600.32

Laspeyres price Index for 1996

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{583.36}{563.92} \times 100 = 103.4472$$

Laspeyres Price Index for 1997

$$\frac{\Sigma p_2 q_0}{\Sigma p_0 q_0} \times 100 = \frac{600.32}{563.92} \times 100 = 106.4548$$

Calculation of Paasche's Index

Model	1995 p_0	1996 p_1	1997 p_2	1996 q_1	1997 q_2	p_0q_1	p_1q_1	p_0q_2	p_2q_2
Vision	17	17.9	18.6	12	12	204	214	204	223
Emperor	13.1	13.6	13.7	9.6	11.2	126	130	147	153
Prince	11.2	11.4	11.9	10.4	12	116	118	134	143
Crown	8.5	8.7	8.8	10.4	12	88	90	102	106
Pearl	5.5	5.6	5.8	14.4	18	79	81	99	104
						613	633	686	729

Paasche's Price Index for 1996

$$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{633}{613} \times 100 = 103$$

$$\frac{\Sigma p_2 q_2}{\Sigma p_0 q_2} \times 100 = \frac{729}{686} \times 100 = 106$$

Fisher's Ideal Index = $\sqrt{L \times P}$

Fisher's price Index for 1996 = $\sqrt{103.447 \times 103} = 103.223$

Fisher's Price Index for 1997 = $\sqrt{106.4548 \times 106} = 106.227$

b. **Calculation of Unweighted Average of Relative Price Index**

Model	1995 P_0	1996 P_1	1997 P_2	$\frac{P_1}{P_0} \times 100$	$\frac{P_2}{P_0} \times 100$
Vision	17	17.9	18.6	105.2941	109.4117
Emperor	13.1	13.6	13.7	103.8167	104.5801
Prince	11.2	11.4	11.9	101.7857	106.2500
Crown	8.5	8.7	8.8	102.3529	103.5294
Pearl	5.5	5.6	5.8	101.8181	105.4545
				515.0677	529.2258

Unweighted average of relative price index for 1996

$$\Sigma\left[\frac{P_1}{P_0} \times 100\right] / n = \frac{515.0677}{5} = 103.0135$$

Unweighted average of relative price index for 1997

$$\Sigma\left[\frac{P_2}{P_0} \times 100\right] / n = \frac{529.2258}{5} = 105.8452$$

Illustration 3

Telly Vision is a player in the television industry. The selling price and the sales figures for the past 3 years are as follows:

	Selling price (in Rs.)			Units sold		
Model	1995	1996	1997	1995	1996	1997
Vision	17,000	17,900	18,600	11,200	12,000	12,000
Emperor	13,100	13,600	13,700	8,000	9,600	11,200
Prince	11,200	11,400	11,900	9,600	10,400	12,000
Crown	8,500	8,700	8,800	11,200	10,400	12,000
Pearl	5,500	5,600	5,800	12,000	14,400	18,000

a. Calculate Laspeyres, Paasche's and Fisher's quantity indices for the year 1996 and 1997 using 1995 as the base year.

b. Calculate unweighted average of relative quantity indices for the year 1996 and 1997 with 1995 as the base year.

Solution

a. **Calculation of Laspeyres Index**

Model	1995 q_0	1996 q_1	1997 q_2	1995	q_0p_0	q_1p_0	q_2p_0
Vision	11.2	12	12.0	17	190.4	204	204
Emperor	8.0	9.6	11.2	13.1	104.8	126	147
Prince	9.6	10.4	12.0	11.2	107.5	116.5	134.4
Crown	11.2	10.4	12.0	8.5	95.20	88.4	102
Pearl	12.0	14.4	18.0	5.5	66.00	79.2	99
					563.9	614.1	686.4

Laspeyres Quantity Index for 1996

$$\frac{\sum q_1p_0}{\sum q_0p_0}\times 100 = \frac{614.1}{563.92}\times 100 = 108.89$$

Laspeyres quantity index for 1997

$$\frac{\sum q_2p_0}{\sum q_0p_0}\times 100 = \frac{686.4}{563.92}\times 100 = 121.71$$

Calculation of Paasche's Quantity Index for 1996 and 1997

Model	1995 q_0	1996 q_1	1997 q_2	1995	1997 p_2	q_0p_1	q_1p_1	q_0p_2	q_2p_2
Vision	11.2	12	12.0	17.9	18.6	200	215	208	223
Emperor	8.0	9.6	11.2	13.6	13.7	109	130	110	153
Prince	9.6	10.4	12.0	11.4	11.9	109	118	114	143
Crown	11.2	10.4	12.0	8.7	8.8	97	90	98	106
Pearl	12.0	14.4	18.0	5.6	5.8	67	81	70	104
						582	635	600	729

Paasche's quantity Index for 1996

$$\frac{\sum q_1p_1}{\sum q_0p_1}\times 100 = \frac{635}{582}\times 100 = 109.10$$

Paasche's quantity Index for 1997

$$\frac{\sum q_2p_2}{\sum q_0p_2}\times 100 = \frac{729}{600}\times 100 = 121.5$$

Fisher's Index = $\sqrt{L\times P}$

Fisher's Quantity Index for 1996 = $\sqrt{108.89\times 109.10}$ =108.99

Fisher's Quantity Index for 1997 = $\sqrt{121.71\times 121.5} = 121.60$

b. **Calculation of Unweighted Average of Relative Quantity Indices**

Model	1995 q_0	1996 q_1	1997 q_2	$\frac{q_1}{q_0}\times 100$	$\frac{q_2}{q_0}\times 100$
Vision	11.2	12	12.0	107.1428	107.1428
Emperor	8.0	9.6	11.2	120	140
Prince	9.6	10.4	12.0	108.3333	125
Crown	11.2	10.4	12.0	92.8571	107.1428
Pearl	12.0	14.4	18.0	120	150
				548.3333	629.2857

Unweighted Average of relative quantity index using 1995 as the base

$$\frac{\sum\left[\frac{q_1}{q_0}\times 100\right]}{n}$$

Index for 1996 = $\frac{548.3333}{5} = 109.667$

Index for 1997 = $\frac{629.2857}{5} = 125.8571$

Illustration 4

Consider the following data:

Commodities	1995		1996	
	Price	Quantity	Price	Quantity
A	4	16	6	13
B	7	10	7	14
C	7	25	10	32
D	3	18	6	22

a. Calculate Laspeyres, Paasche's, Fisher's, Marshall-Edgeworth and Dorbish and Bowley's Price indices .

b. Verify the Fisher's index satisfies time reversal and factor reversal test.

Solution

p_0	q_0	p_1	q_1	$p_0 q_0$	$p_0 q_1$	$p_1 q_0$	$p_1 q_1$
4	16	6	13	64	52	96	78
7	10	7	14	70	98	70	98
7	25	10	32	175	224	250	320
3	18	6	22	54	66	108	132
				363	440	524	628

Laspeyres price Index

$$\frac{\sum p_1 q_0}{\sum p_0 q_0}\times 100 = \frac{524}{363}\times 100 = 144.3526$$

Paasche's Price Index

$$\frac{\sum p_1 q_1}{\sum p_0 q_1}\times 100 = \frac{628}{440}\times 100 = 142.7273$$

Fisher's Ideal Index = $\sqrt{L \times P}$

$$= \sqrt{\frac{524}{363} \times \frac{628}{440}} \times 100 = 143.5376$$

Marshall-Edgeworth's Index

$$\frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0} \times 100$$

$$\frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

$$\frac{524 + 628}{363 + 440} \times 100 = 143.462$$

Dorbish and Bowley's Price Index

$$\frac{L + P}{2} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}}{2}$$

$$= 144.3526+142.7273/2 = 143.53995$$

The time reversal test, states that $I_{0,1} \times I_{1,0} = 1$

$$I_{0,1} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} = \sqrt{\frac{524}{363} \times \frac{628}{440}}$$

$$I_{1,0} = = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = \sqrt{\frac{440}{628} \times \frac{363}{524}}$$

$$I_{0,1,} \times I_{1,0} = \sqrt{\frac{524}{363} \times \frac{628}{440}} \quad x \quad \sqrt{\frac{440}{628} \times \frac{363}{524}} = 1.$$

Therefore, Fisher's index satisfies time reversal test.

Fisher's Quantity Index

$$Q_{0,1} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} = \sqrt{\frac{440}{363} \times \frac{628}{524}}$$

$$P_{0,1} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} = \sqrt{\frac{524}{363} \times \frac{628}{440}}$$

Factor's Reversal Test

Fisher's price index x Fisher's quantity index

$$= \sqrt{\frac{524}{363} \times \frac{628}{440}} \quad x \quad \sqrt{\frac{440}{363} \times \frac{628}{524}}$$

$$= \frac{628}{363} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Therefore, Fisher's index satisfies factor reversal test.

Illustration 5

The following table contains information from the raw material purchase records of a tire manufacturer for the years 1981, 1982, and 1983.

Material	Average Annual Purchase Price/ton			Value '000
	1981	1982	1983	1983
Butadiene (B)	17	15	11	50
Styrene (S)	85	89	95	210
Rayon Cord (R)	348	358	331	1640
Carbon Black (C)	62	58	67	630
Sodium Pyrophosphate (SP)	49	567	67	90

Calculate a weighted average of relative price index for each of those 3 years, using 1983 for weighting and for the base year.

Solution

	P_1	P_2	P_0	$p_0 q_0$	$\frac{P_1}{P_0}\times 100$	$\frac{P_2}{P_0}\times 100$		
1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) x (6)	(5) x (7)
B	17	15	11	50	155	136	7,750	6,800
S	85	89	95	210	90	94	18,900	19,740
R	348	358	331	1640	105	108	1,72,200	1,77,120
C	62	58	67	630	93	87	58,590	54,810
SP	49	56	67	90	73	84	6,570	7,560
				2,620			2,64,010	2,66,030

Index for 1981 = $\frac{2,64,010}{2,620} = 100.78$

Index for 1982 = $\frac{2,66,030}{2,620} = 101.54$

Index for 1983 = 100.

Illustration 6

The information described the unit sales of a bicycle shop for 3 years:

	Number Sold			Price
	1981	1982	1983	1981
A	45	48	56	89
B	64	67	71	104
C	28	35	27	138
D	21	16	28	245

Calculate the weighted average of relative quantity indices, using the price and quantities from 1981 to compute the value weights, with 1981 as the base year.

Solution

	Number Sold			Price			
	1981 q_0	1982 q_1	1983 q_2	1981 p_0	q_0p_0	q_1p_0	q_2p_0
A	45	48	56	89	4,005	4,272	4,984
B	64	67	71	104	6,656	6,938	7,364
C	28	35	27	138	3,864	4,830	3,726
D	21	16	28	245	5,145	3,920	6,860
					19,670	19,990	22,954

Weighted Average of Relative Quantity Indices = $\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100$

Index for 1982 = $\frac{19,900}{19,670} \times 100 = 101.63$

Index for 1983 = $\frac{22954}{19670} \times 100 = 116.70$

Illustration 7

You are given the following price indices for items A and B:

Year	Price Relatives for A	Unweighted Average of Relative Price Index for A and B
1992	100	100
1993	120	130
1994	135	150

Calculate weighted aggregates price index for A and B giving a weightage to A, which is three times the weightage you give to B with the base year 1992.

Solution

Let X be the percentage price relative for B in 1993, then

$$\frac{\frac{120}{100} \times 100 + \frac{X}{100} \times 100}{2} = 130$$

i.e., 120 + X = 260

X = 140

Similarly, if Y is the percentage price relative for B in 1994

$$\frac{\frac{135}{100} \times 100 + \frac{Y}{100} \times 100}{2} = 150$$

i.e., 135 + Y = 300

Y = 165

Calculation of Weighted Average Price Index

Year	A	B	Index
1992	100	100	100
1993	120	140	$\frac{120 \times 3 + 140 \times 1}{400} \times 100 = 125$
1994	135	165	$\frac{135 \times 3 + 165 \times 1}{400} \times 100 = 142.5$

Illustration 8

A company produces three products viz., erasers (E), pencils sharpeners(PS) and ball point pens(BP). The table below gives prices and units sold (Qty) in each of the year 2002, 2003 and 2004.

Products	2002		2003		2004	
	Price	Qty.	Price	Qty.	Price	Qty
E	0.50	30	0.75	30	1	35
SP	2.00	25	4.00	20	3	25
BP	3.00	20	3.50	25	3.50	20

Using 2002 as the base, Calculate

a. The unweighted aggregates price index and quantity index for 2004.

b. The weighted aggregates index for 2004, by using both Laspryres method and Paasches method.

c. The unweighted average of relative price index and unweighted average of relative quantity index for 2004.

Solution

a. Unweighted aggregates price index for 2004 using 2002 as base

$$= \frac{1+3+3.50}{0.50+2+3}$$

$$= 136.36$$

Unweighted aggregates quantity index for 2004 using 2002 as base

$$= \frac{35+25+20}{30+25+20} \times 100$$

$$= 106.67$$

b. Weighted aggregates price index for 2004 using Laspeyres method

$$= \frac{(1\times 30)+(3\times 25)+(3.50\times 20)}{(0.50\times 30)+(2\times 25)+(3\times 20)} \times 100$$

$$= \frac{175}{125} \times 100$$

$$= 140$$

Weighted aggregates price index for 2004 using Paasches method

$$= \frac{(1\times 35)+(3\times 25)+(3.50\times 20)}{(0.50\times 35)+(2\times 25)+(3\times 20)} \times 100$$

$$= \frac{180}{127.5} \times 100$$

$$= 141.18$$

c. Unweighted average of relative price index for 2004

$$= \frac{(1/0.50\times 100)+(3/2\times 100)+(3.5/3\times 100)}{3}$$

$$= \frac{200+150+116.67}{3} = 155.56$$

Unweighted average of relative quantity index for 2004

$$= \frac{(35/30\times 100)+(25/25\times 100)+(20/20\times 100)}{3}$$

$$= \frac{116.67+100+100}{3} = 105.56$$

Illustration 9

The Laspeyres and Fisher's ideal index for a quantity are 103.28 and 108.91 respectively. Is it possible to find out Paasches's index from this data? If yes, then what is your answer. If no, why?

Solution

We know that Fisher's Ideal Index is Geometric mean of Laspeyres and Paasche's Index

$$\text{Fisher's Index} = = \sqrt{L \times P}$$

$$\text{Paasche's Index} = \frac{(\text{Fisher's Index})^2}{\text{Laspeyres Index}}$$

$$= \frac{(108.91)^2}{103.28}$$

$$= 114.85$$

Illustration 10

From the following average price of the group of commodities given in rupees per unit, find chain base index number with 1994 as the base year

Group	1994	1995	1996	1997	1998
I	1	1.5	2	2.5	3
II	4	5	6	7.5	9
III	2	2.5	9	5	6

Solution

Calculation of Chain Base Index Number with 1994 as Base Year

Group	1994		1995		1996		1997		1998	
	Price	Link Relative	Price	Link Relative	Price	Link Relative	Price	Link Relative	Price	Link Relative
I	1	100	1.5	150	2	133.3	2.5	125	3	120
II	4	100	5	125	6	120.0	7.5	125	9	120
III	2	100	2.5	125	4	160.0	5	125	6	120
Total		300		400		413.3		375		360
Avg. link relative		100		133.3		137.77		125		120
Chain index		100		133.3		183.69		229.61		275.53

$$\text{Chain Index for 1995} = \frac{133.33}{100} \times 100 = 133.33$$

$$\text{Chain Index for 1996} = \frac{133.33}{100} \times 137.77 = 183.69$$

$$\text{Chain Index for 1997} = \frac{183.69}{100} \times 125 = 229.61$$

$$\text{Chain Index for 1998} = \frac{229.61}{100} \times 120 = 275.53$$

$$\text{Chain index for 1998} = \frac{229.61}{100} \times 120 = 275.53$$

SUMMARY

- Index numbers are one of the most widely used statistical indicators. Generally they are used to indicate the state of the economy, index numbers are aptly called 'barometers of economic activity'. An index number is a statistical measure designed to show changes in a variable or a group or related variables with respect to time, geographic location or other characteristics such as income, profession, etc. It is calculated as a ratio of the current value to a base value and is expressed as a percentage. The index number for the base year is always 100. Index numbers are very useful in revealing a general trend of the phenomenon under study, in policy-making, in determining purchasing power of the rupee and in deflating time series data (i.e., adjusting the original data to reflect reality).

- There are three types of principle indices:
  i. Price index (compares changes in price from one period to another).
  ii. Quantity index (measures change in quantity from one period to another).
  iii. Value index (combines price and quantity changes to present a more spatial comparison).
- There are two approaches for constructing an index number, namely the aggregates method and average of relatives method. The index constructed in either of these methods could be an unweighted index or a weighted index.
  – An unweighted aggregate index is calculated by totaling the current year/given year's elements and then dividing them by the sum of the same elements during the base period.
  – In weighted aggregate index, weights are assigned to the various items according to their significance. The important methods of assigning weights to an index are:
    1. Laspeyres method (uses the quantities consumed during the base period in computing the index number).
    2. Paasche's method (uses quantity measures for the current period rather than the base period).
    3. Fixed weight aggregates method (the weights used are neither from the base period nor from the current period, but from a representative period).
    4. Fisher's ideal index number (is the geometric mean of the Laspeyers and Paasche's indices).
    5. Marshall Edgeworth method (uses both current year as well as base year prices and quantities).
  – In unweighted average of relatives method, when a price index is constructed, all price relatives (quantity relatives in case of quantity index) are to be obtained for all items included in the index and then the average of these price relatives is computed.
  – For weighted average of relatives method, the value of each element in the composite is used as the weight. The values could be from the base year, current or any fixed year. It can be of fixed-based or chain-based.
- The consistency of index numbers can be tested using the time reversal test, factor reversal test and the circular test.

Chapter XII

Time Series Analysis

After reading this chapter, you will be conversant with:

- Time Series Analysis
- Procedure for Fitting a Straight Line
- Components of Time Series Analysis (Secular Trend, Cyclical Variation, Seasonal Variation (with ratio of moving averages) and Irregular Variation)
- Comprehensive Illustration
- Time Series Analysis in Forecasting
- Additional Illustrations

Introduction

One of the prerequisites for a firm to succeed in a fiercely competitive environment is to gain insights into the future, so that they can position the firm in a state of preparedness, to meet the challenges that future may entail. This would not appear to be such a great problem, if managers are certain about the course of action future would take. That is, the element of uncertainty makes deciphering the future difficult and interesting. To achieve this objective, the managers in the course of decision-making use various models and techniques. One such technique used in almost all the spheres of business management is "Forecasting".

The marketing department has to make short-term as well as long-term sales forecast; the production department has to forecast the inventory levels and the finance department has to forecast the level of working capital required keeping in view the requirements of other departments.

It is also true that the forecasts, which are obtained, are not accurate to the last detail. Under these conditions, it can be said that the success of a firm depends on how well the management is able to identify a suitable technique and employ it to process the available data to reach a valid conclusion.

The first time we came across the forecasting concept was in Interpolation and Extrapolation and to a great extent in regression analysis. In fact, the time series analysis, which we will look at in this chapter, is also called extrapolation method and it also utilizes the equations of best fit, which we have seen in the regression analysis albeit with some modifications. Therefore, by utilizing this technique, we can cope with the degree of uncertainty at least to some extent.

TIME SERIES ANALYSIS

A time series is an arrangement of statistical data in a chronological order. Time series occupies an important and significant place in business and economics. One of the important tasks before economists and businessmen is to make estimates for the future. For example, a businessman is interested in finding out the likely sales for the year 2006 so that he can adjust the production and avoid locking of working capital in unsold stock. Similarly, an economist is interested in estimating the likely increase in population so as to make plans for food supply, creation of employment opportunities and other social cost. For making this estimate, one need to collect the statistical data at successive intervals of time. Time series is the set of observation of numerical data at different point of time, in which time is the most important factor and the variables are related to time. So, in the Time Series Analysis, we try to identify and determine the pattern of changes in the data collected over regular intervals of time. The data collected can be at a periodical interval of days, weeks, months and years. After identifying the patterns, we project them into future to get an estimate of the variable under consideration. Do the changes or variations observed in different time series be the same or different? Necessarily, they should be different as the independent variable (X) on the axis, which influences the dependent variable differs from one time series to another.

Definition of Time Series:

According to Patterson "A time series consists of statistical data which are collected, recorded, observed over successive increments."

According to Morris Hamburg "A time series is a set of statiscal observations arranged in chronological order."

According to Ya-lun-Chou, "A time series may be defined as a collection of magnitudes belonging to different periods, of some variables or composite of variables, such as production of steel, per capital income, gross national product, price of tobacco, or index of industrial production".

Steps in the Analysis of a Time Series

- Identifying or determining the various forces or influences whose interaction produces variations in the time series.
- Isolating, studying, analyzing and measuring them independently, i.e., by holding other things constant.

Components of a Time Series

Broadly, the variations observed in a time series can be classified as (i) the secular trend – is a change that occurs due to general tendency of the data to increase or decrease, (ii) the cyclical fluctuation – are the changes that occur as a result of booms and depressions, (iii) the seasonal variation – variations/changes that occur due to change in climate, weather conditions etc., and finally, (iv) the irregular variations are the changes that occur due to unpredicted forces such floods, famines earthquakes etc. They are collectively called components of time series.

Editing of a Time Series

Editing of a Time Series is necessary in the following circumstances:

- If the data is available on the monthly basis it should be adjusted because all the months do not have same number of days. This is done by calculating the daily average of a particular month's data and multiplying it by 365/12 i.e., average number of days in a month. In other words, it is the product of daily average and the average number of days in a month.
- Adjustment to population changes is necessary where the variable gets affected by population. The example of such variable may be national income. By dividing national income by the no. of working persons in the country we get the per capita income.
- Adjustment to price changes is necessary to derive real value changes from the current value series.
- The data collected for the purpose of analysis must be homogenous and the data analyzed should be capable of comparability to develop valid conclusions.

Utility of Time Series Analysis/Need for Conducting Time Series Analysis

Time series analysis is of great significance not only to the businessmen and economists, but also to scientists, astronomists, research workers etc., for the following reasons:

- It helps in understanding the changes that take place in past behavior by observing the data over a period of time. It also helps in predicting the future behavior.
- It helps in planning the future operations. If the occurrence of any event over a long period of time is established within the limits, then prediction of probable future is possible.
- It helps in evaluating current accomplishment by comparing the actual performance with the predetermined performance, then analyzing the cause of variations.
- It facilitates comparison of data over a period of time and helps in drawing the conclusions there from.

PROCEDURE FOR FITTING A STRAIGHT LINE

Before actually going into the detailed study of different components of time series, let us discuss the procedure for fitting a trend line through "Linear and Multiple Regression Analysis". It forms the basis for components of a time series.

The following is the procedure for fitting a straight line through "Linear and M\ultiple Regression Analysis":

a. The equation of a straight line is given by Y = a + bX, where X and Y are independent and dependent variables respectively, "a" is the Y-intercept (the value of Y when X is equal to zero. That is Y = a + b.0 = a + 0 = a) and finally "b" is the slope of the straight line.

b. For the given values of X and Y, we run regression analysis to get the estimating equation, which is of the form $\hat{Y} = a + bx$, where "x" represents the deviations taken from "X". At this point, it is worthwhile to note that the general trend of the given time series can be expressed in terms of any number of straight lines. But what we are concerned with, is the straight line of the best fit. That is, the deviations of the actual series and the estimated points, when squared and added should give a value, which is minimum if we repeat the process for other straight lines obtained from the given data. This can be achieved if we employ the equations for "a" and "b" given below:

$$b = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2},$$

$$a = \overline{Y} - b\overline{X}$$

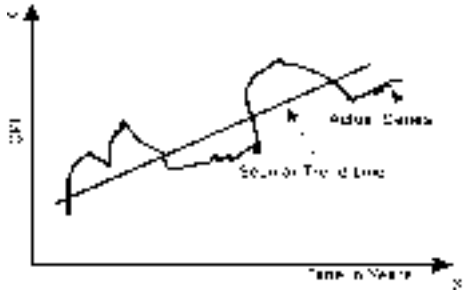
where "a" and "b" are the Y-intercept and the slope of the estimating equation, "n" is the number of data points in the series, $\overline{X}$ and $\overline{Y}$ are the mean values of X and Y data points.

COMPONENTS OF A TIME SERIES ANALYSIS

Secular Trend

Under this type of variation, we look at the long-term behavior of the variable. This is important in the sense that one should be able to conclude whether the value of the variable has been increasing or decreasing over the years, keeping aside the variations observed in the individual years that constitute the series. An appropriate example of secular trend would be the variation in the Consumer Price Index (CPI) observed over a period of say 20 years. The Secular Trend can be represented graphically as shown in the figure 1. In the figure, the secular trend is shown as a straight line, with an upward slope, while the actual time series is shown as a curve moving towards and away from the trend line.

Figure 1



The next logical question would be, apart from deciding the long-term trend of the variable, what are the other advantages of studying the secular trend? They can be listed as:

i. By studying the secular trend, one can examine whether a policy implemented has yielded the necessary results or not and what would be its future impact.

For example, if a decision is taken to introduce multiple quality checks for a product, it is possible to use the secular trend in the past to verify whether the decision resulted in a significant reduction in poor quality products getting into the market. Similarly, if a department is constituted to follow the receivables of a company, it is possible to see whether this measure has resulted in any improvement in the recovery of receivables.

ii. After studying the secular trend, one can project it into the future to get an estimation (future value) of the variable. This helps us to take necessary measures in time.

A secular trend gives us an understanding about the behavior of a variable. Hence, it will be possible to project its value to a future date to know what the value is likely to be. For example, a company which proposes to develop a medium-term plan can have a sales forecast based on the historical trend.

iii. Under certain circumstances, it is required to examine the time series for other variation components only. Therefore, by studying the secular trend one can separate it, which in turn facilitates the study of other components present in the series.

While studying the trends in inflation, it is possible to study the impact of seasonal variations for which separation of the impact of secular trend is necessary.

In the above figure, we have seen that the secular trend is a straight line. Do all the trend lines need to be linear? No, it is not mandatory that all the trend lines are to be linear in nature. It depends on the phenomena which we are trying to explain. This is because, some relationships are amenable to be dealt by using linear (straight line) models, while the characteristics of some others can be better brought out and examined if we use a curvilinear model. For example, the pollutants in the environment do not increase in a linear fashion. In this case, a curvilinear model may be more appropriate as compared to a linear model. Can we fit the trend, whether it be straight line or curvilinear only by inspecting the points visually or follow a scientific procedure? Yes, we can fit a trend observing the points visually. However, the problem with this method is that different people tend to have their own bias giving rise to ambiguity. Therefore, to get a best fit, we employ the least squares method, which we have first seen in Linear and Multiple regression. Let us recollect how we have fitted a straight line.

a. The equation of a straight line is given by $Y = a + bX$, where X and Y are independent and dependent variables respectively, "a" is the Y-intercept (the value of Y when X is equal to zero. That is $Y = a + b.0 = a + 0 = a$) and finally "b" is the slope of the straight line.

b. For the given values of X and Y, we run regression analysis to get the estimating equation, which is of the form $\hat{Y} = a + bX$. At this point, it is worthwhile to note that the general trend of the given time series can be expressed in terms of any number of straight lines. But what we are concerned with, is the straight line of the best fit. That is, the deviations of the actual series and the estimated points, when squared and added should give a value, which is minimum if we repeat the process for other straight lines obtained from the given data. This can be achieved if we employ the equations for "a" and "b" given below:

$$b = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2},$$

$$a = \overline{Y} - b\overline{X}$$

where "a" and "b" are the Y-intercept and the slope of the estimating equation, "n" is the number of data points in the series, $\overline{X}$ and $\overline{Y}$ are the mean values of X and Y data points.

Since in time series, the independent variable X in most of the cases happens to represent years (2000, 2001, 2002,.....), our computational part will become tedious if we use it as it is. In order to make it easier, we take the mean of the X data points and subtract it from actual X values and represent that column as "x". This process is referred to as "Coding or Translating time". Therefore, when there are **odd number of data points**, the middle point will become zero. When there are **even number of data points**, the mean will be like 2001.5. This requires that after subtracting the mean from the actual data points, we multiply the resultant values with 2 and denote that column as "x". In other words, "x" denotes the coded time in terms of half year intervals. The other advantage of this process is that, the value of the mean happens to be zero. Incorporating these changes in the equations for "a" and "b", we get

$$b = \frac{\sum xY}{\sum x^2} \text{ and } a = \overline{Y}$$

We now look at a couple of examples to understand the concept of coding.

Illustration 1

The number of PCs sold from 1996 to 2002 are:

Years	Number of PCs (in thousands)
1996	51
1997	54
1998	59
1999	61
2000	63
2001	65
2002	69

Fit a trend line for this data.

Solution

We observe that the number of data points is odd. The working is shown below:

	Years (X)	Number of PCs sold (Y)	$x = X - \overline{X}$	x.Y	x^2
	(1)	(2)	(3)	(4)	(5)
	1996	51	–3	–153	9
	1997	54	–2	–108	4
	1998	59	–1	–59	1
	1999	61	0	0	0
	2000	63	1	63	1
	2001	65	2	130	4
	2002	69	3	207	9
Total	13993	422		80	28

$$\overline{X} = \frac{13993}{7} = 1999$$

We have $b = \frac{\sum xY}{\sum x^2} = \frac{80}{28} = 2.86$, and

$$a = \bar{Y} = \frac{422}{7} = 60.29$$

Therefore, the regression equation describing the secular trend is given by

$$\hat{Y} = 60.29 + 2.86x$$

Illustration 2

The data for this problem is same as above except that in the year 1995 the sale of PCs was 48,000. We fit a secular trend for this data.

Solution

The number of data points after including the sales in 1995 is even. The calculations are shown below:

Years (X)	Number of PCs sold (Y)	$Z = X - \bar{X}$	$x = 2.Z$	(2).(4)	x^2
(1)	(2)	(3)	(4)	(5)	(6)
1995	48	–3.5	–7	–336	49
1996	51	–2.5	–5	–255	25
1997	54	–1.5	–3	–162	9
1998	59	– 0.5	–1	–59	1
1999	61	0.5	1	61	1
2000	63	1.5	3	189	9
2001	65	2.5	5	325	25
2002	69	3.5	7	483	49
Total 15988	470			246	168

$$\bar{X} = \frac{15988}{8} = 1998.5$$

We have $b = \frac{\sum xY}{\sum x^2} = \frac{246}{168} = 1.46$ and

$$a = \bar{Y} = \frac{470}{8} = 58.75$$

Therefore, the regression equation describing the secular trend is given by

$$\hat{Y} = 58.75 + 1.46x$$

We have seen the method to fit a straight line to the linear trend in case of both even and odd number of data points. If we want to estimate the number of PCs that would be sold in 2006, we proceed as follows. The year 2002 should be translated as 2006 – 1999 = 7 (we took 1999 as the mean year in the previous example) and substituted in the estimating equation $\hat{Y} = 60.29 + 2.86x$. That will be

$$\hat{Y} = 60.29 + 2.86(7) = 80.31.$$

That is, the demand for the PCs in the year 2006 is expected to be 80,310.

The estimate when we have even number of data points would be 2006 – 1998.5 = 7.5 years. Multiplying this by 2, we have 7.5 x 2 = 15 half-year intervals. Substituting this in the estimating equation, we have

$$\hat{Y} = 58.75 + 1.46(15) = 80.65$$

That is, the demand in the year 2006 will be 80,650.

Now we look at best fit, for a curvilinear trend. The curvilinear trend can be represented by a parabola whose equation is $Y = a + bx + cx^2$. The corresponding estimating equation is given by $\hat{Y} = a + bx + cx^2$ ("x" represents the coded time variable). As we have seen in the case of linear trend, the values of the constants a, b, and c for the parabola of the best fit are given by solving these equations.

$$\sum Y = an + c\sum x^2$$

$$\sum x^2 Y = a\sum x^2 + c\sum x^4 \text{ and } b = \frac{\sum xY}{\sum x^2}$$

Now, we look at an example to obtain a best fit for a parabola.

Illustration 3

The number of color televisions sold during the period 1998 to 2002 is as shown below:

Year (X)	Number of Sets Sold (in thousands) (Y)
1998	55
1999	58
2000	63
2001	65
2002	67

Solution

For this data, we fit a curvilinear trend. The calculations pertaining to this example are shown below:

	Years (X)	Number of Sets Sold (Y) (in thousands)	$x = X - \bar{X}$	x^2	x^4	xY	x^2Y
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	1998	55	–2	4	16	–110	220
	1999	58	–1	1	1	–58	58
	2000	63	0	0	0	0	0
	2001	65	1	1	1	65	65
	2002	67	2	4	16	134	268
Total	10,000	308		10	34	31	611

$$\text{Mean of } X = \frac{10,000}{5} = 2,000$$

We have $\sum Y = an + c\sum x^2$

$$\sum x^2 Y = a\sum x^2 + c\sum x^4 \text{ and } b = \frac{\sum xY}{\sum x^2}$$

Substituting the values in these equations, we have

$308 = 5a + 10c$ ... (i)

$611 = 10a + 34c$ ... (ii)

$b = \frac{31}{10} = 3.1$ ... (iii)

Multiplying (i) by 2, we have $616 = 10a + 20c$ …(iv)

We solve (iv) and (ii) $616 = 10a + 20c$

$611 = 10a + 34c$

On subtracting, we have $5 = -14c$

$c = -5/14$

$= -0.357$

Substituting the value of $c = -0.357$ in (i), we have

$308 = 5a + 10(-0.357)$

$308 = 5a - 3.57$

$308 + 3.57 = 5a$

That is; $5a = 311.57$

$a = 311.57/5 = 62.314$

On substituting the values of a, b and c in the equation of the parabola, we have

$\hat{Y} = 62.314 + 3.1x - 0.357x^2$

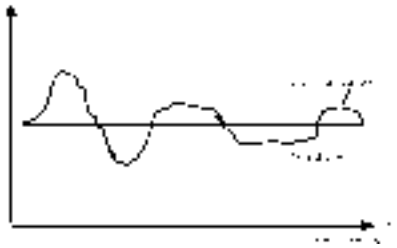
One of the points we have to remember while using a relationship obtained either by a linear trend or a curvilinear trend is that the relationship cannot be used across any range of time for forecasting. For example, the demand for color televisions may not increase at the pace indicated by the equation of the parabola. We are aware of the fact that the life cycle of any product does not extend to an infinite period. Due to advancement of technology, one can always come up with an improvised and a more economical alternative. Also, the price of a substitute may also decrease causing a fall in the demand for the original product. Therefore, while employing the trends for forecasting purposes, one should be more careful and avoid making a sweeping statement solely based on the forecast.

Cyclical Variation

By cyclical variations, we refer to the long-term movement of the variable about the trend line. Therefore, does the movement of the actual series about a trend line observed in figure (1) suggest a pattern of cyclical variation? Yes, it does. A more appropriate example of cyclical variation would be the pattern of business cycle. The whole cycle encompassing the phenomena of the business activity reaching a high, its gradual slow down, a depression and then the recovery can be observed. But one of the points to which we have to pay attention is that the business and the economic cycles may not be periodic in nature. That is, in two different business cycles, at equal intervals of time, the pattern observed may be different.

Generally, the behavior of the variable is considered to be cyclic, only if the movements recur after a period of more than one year. The cyclical variation is shown in the figure 2.

Figure 2



Source: Adapted from Gupta S.P., Statistical Methods.

In the figure, we observe the component of the time series moving towards and away from the secular trend.

RESIDUAL METHOD

We know that a time series consisting of annual data for longer periods is depicted by trend lines. This facilitates us to isolate the component of secular trend variation from the series and examine it for cyclical, seasonal and irregular components. In this part, we will look at "Residual Method", by which one can isolate the cyclical variation component. Further, this method can be bifurcated into two measures: Percent of Trend and Relative Cyclical Residual measures. Both these measures are expressed in terms of percentage. We look at each of them.

Percent of Trend Measure

When the ratio of actual values (Y) and the corresponding estimated values ($\hat{Y}$) is multiplied by 100, we are expressing the cyclical variation component as a percent of trend. Mathematically, we express it as

$$((Y/\hat{Y}) \times 100)$$

When percent of trends are calculated and plotted on a graph, we can observe the variations from the trend line. Let us look at an example, which explains this method.

Relative Cyclical Residual Measure

In this measure, we take the ratio of the difference between the Y and the corresponding. $\hat{Y}$ values (that is, $Y - \hat{Y}$), and the $\hat{Y}$ values. To express these values in terms of percentage we multiply them by 100. In other words, the percentage deviation from the trend is found for all the values in the series. Mathematically, this is expressed as:

$$\frac{Y - \hat{Y}}{\hat{Y}} \times 100$$

In the above example, the values of Relative Cyclical Residual are obtained when 100 is subtracted from the values given in column (7). The respective values are shown in column (8). The plot of Relative Cyclical Residual is shown below. A value of –5.81 indicates that the number of cartons of cereal sold by the Departmental Store has shown a decrease of 5.81% from its previous level.

Illustration 4

The number of cartons of cereal sold by a Departmental store is shown below. Calculate the percent of trend measure and relative cyclical residual measure and plot them on the graph.

Years	Number of Cartons Sold (Y)
1996	18
1997	21
1998	23
1999	23
2000	25
2001	26
2002	27

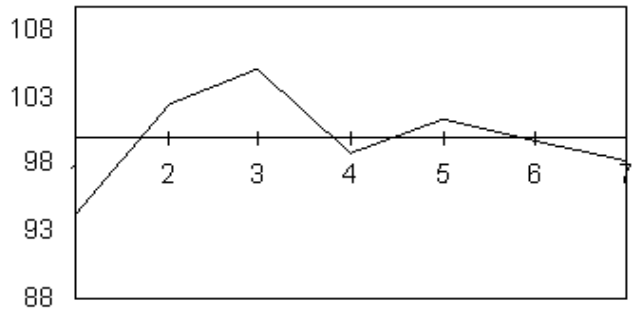
Solution

	Years (X)	Y	x	xY	x^2	$\hat{Y}$ (a + bx)	Percent Trend	Relative Cyclical Residual
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	1996	18	–3	– 54	9	19.11	94.19	–5.81
	1997	21	–2	– 42	4	20.50	102.43	+2.43
	1998	23	–1	– 23	1	21.89	105.07	+5.07
	1999	23	0	0	0	23.28	98.79	–1.21
	2000	25	1	25	1	24.68	101.29	+1.29
	2001	26	2	52	4	26.07	99.73	–0.27
	2002	27	3	81	9	27.46	98.32	–1.68
Sum	13993	163		39	28			

$$\overline{X} = \frac{13993}{7} = 1999$$

$$\overline{Y} = a = \frac{\sum Y}{n} = \frac{163}{7} = 23.29$$

$$a = 23.29, \qquad b = \frac{\sum xY}{\sum x^2} = \frac{39}{28} = 1.3929$$

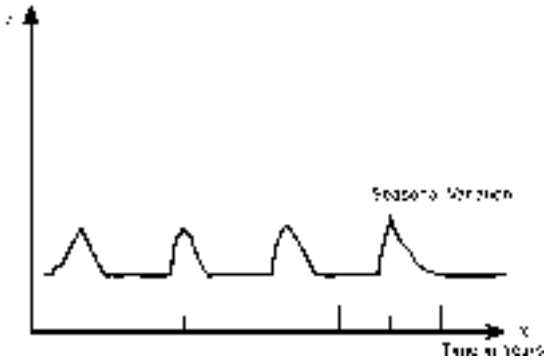


In the figure, we observe the components of the cyclical variation after eliminating the secular trend.

Seasonal Variation

Under this variation, we observe that the variable under consideration shows a similar pattern during certain months of the successive years. An example of seasonal variation would be an increase in water-borne diseases during rainy season. That is, this phenomenon seems to be repeating itself every year in a regular fashion. Usually, the time period over which this variation is considered can consist of days, weeks, months and at the most one year. The seasonal variation is depicted in the figure below. We can observe the uniformity in the pattern during every second quarter of the year.

Figure 4



The study of trends has a positive impact on the overall analysis and facilitates the process of drawing more meaningful conclusions. The analysis is useful because

i. It helps us to establish the pattern of changes during the same period in different years and examine the reasons for the same.

ii. It helps us to project the past patterns into future. The projections from the seasonal variations are helpful in formulating short-term objectives of a firm. This is similar to the projection of the secular trend into the future for achieving the long-term goals.

iii. It helps the identification of the seasonal trend in a time series to isolate it and study the impact of other components of variation in the series.

METHOD TO IDENTIFY THE COMPONENT OF SEASONAL VARIATION IN A TIME SERIES

This technique is called Ratio to Moving Average Method. In this technique, we construct an index which has a base of 100. The magnitude of seasonal variations is measured by the individual deviations from the base of 100. The six steps employed in the construction of the seasonal index are explained with the help of an example.

MOVING TOTALS AND MOVING AVERAGES

Before we look at the six steps constituting the construction process, we look at moving totals and moving averages. We will come across these steps in the calculation of the seasonal index. Consider 4, 6, 2, 7, 9, 8 and 4 which is a set (collection) of numbers. For this set of numbers, if we compute the moving totals of order 3, they will be like 4 + 6 + 2, 6 + 2 + 7, 2 + 7 + 9, 7 + 9 + 8 and 9 + 8 + 4. From this, the meaning of order should be clear. It refers to the number of elements, we ought to treat as a single group every time we calculate the moving total. We also note that in each subsequent calculation, we exclude the first number and include the number which comes immediately after the last number of the first group. That is the moving total for the second group includes 7, which immediately succeeds 2. This is why precisely, it is called moving total. The moving totals for the above set of numbers will be 12, 15, 18, 24, and 21.

Then the **moving averages** for this set of numbers refer to arithmetic mean of the above moving totals. That is, each value in the set of moving totals should be divided by the order. Therefore, the moving averages will be 4, 5, 6, 8, and 7 respectively. Now observe the relative position of the moving average with respect to the original data.

Original data	4,	6,	2,	7,	9,	8,	4
Moving average		4,	5,	6,	8,	7	

The objective of expressing the data in this form is to point out that the values in the moving averages row is the mean of the three numbers immediately above it. In other words, they are already centered. We do not have to center them further. Now is this the case if we have even number of data points? No. Whenever, even number of data points are present we have to center the moving average. To understand this, consider the same set of numbers we considered above except that we leave out the last number. That is, the set now consists of 4, 6, 2, 7, 9, 8. The mid point for this set is between the third and the fourth values. Now we compare the moving averages of order four and the original data.

Original data	4,	6,	2,	7,	9,	8
Moving average		4.75,		6,	6.5	

We observe that 4.75 falls between the second and the third data points. In this case, centering would be to associate the moving average with either the second or the third data point. For the given number set we can associate the moving average with the third data point. This we do by taking the average of the moving averages 4.75 and 6, which is 5.375. Therefore, 5.375 and 6.25 are associated with 2 and 7 respectively.

The centered averages then can be shown as below:

Original data	4,	6,		2,		7,		9,	8,
Moving average			4.75		6,		6.5		
Centered moving average				5.375		6.25			

Thus, whenever even number of data points are given the moving averages are required to be centered. In the case of odd number of data points, the moving averages are already centered and hence no longer need to be centered further.

Primarily, the concept of moving averages is used to smoothen out the fluctuations inherent in the time series data.

The order of the moving average can be expressed in terms of days, weeks, months and years depending on the context.

Now, we take up an example to understand how the component of seasonal variation is computed.

Illustration 5

Given below is the data, regarding the amount of the foodgrains exported during the past four years on a quarterly basis in Rs. billions. For this data, determine the seasonal component.

Years	Quarter I	Quarter II	Quarter III	Quarter IV
1998	1	2	2	1
1999	3	2	4	3
2000	6	7	8	8
2001	4	5	5	6

Step 1: To begin with, we compute the moving total by considering the values in the four quarters of all the years. Observe that we have even number of data points. After the calculation of the sum, we place them at the middle of the values from which they were calculated. This can be seen in column three of the table 1. We observe that 6 (= 1 + 2 + 2 + 1). This we do, as the moving total is associated with the centermost point among the given values. Also, note that as we move downwards, we exclude the value of the first quarter and include the value which comes next. That is, the value in the first quarter of the next year, which in this case happens to be 3. Therefore, the next moving total would be 8 (= 2 + 2 + 1 + 3).

Step 2: In this step, we calculate the average of the moving totals calculated in the step 1. Thus, 6/4 = 1.50 and 8/4 = 2. Similarly, we calculate the average for the rest. Since we are considering four quarters as a single group, we divide the moving total by four.

Step 3: Since the number of data points is even, we need to center the moving averages. We take the average of first two moving averages and associate it with the third data point, that is 12.1. Similarly, we center the rest of the moving averages, which can be seen in the column (5) of table 1. As studied above, this process smoothens out the fluctuations in the time series data. The data at this point would contain trend and cyclical components of variation.

Step 4: In this step, we calculate the percentage of the actual data to the centered moving average values. This will be (2/1.75) x 100 = 114.30. By this step, we recover the seasonal components for the respective quarters, which has been smoothened when we calculated the moving averages. These percentages are shown in the column (6) of table 1.

Table 1

Year	Quarter	Amount Exported	Four Quarter Moving Total	Four Quarter Moving Average	Centered Moving Average	Percentage of Actual to CMA
	(1)	(2)	(3)	(4)	(5)	(6)
1998	I	1				
			–			
	II	2				
			6	1.50		
	III	2			1.75	114.30
			8	2.00		
	IV	1			2.00	50.00
			8	2.00		
1999	I	3			2.25	133.30
			10	2.50		
	II	2			2.75	72.73
			12	3.00		
	III	4			3.38	118.34
			15	3.75		
	IV	3			4.38	68.49
			20	5.00		
2000	I	6			5.50	109.10
			24	6.00		
	II	7			6.63	105.60
			29	7.25		
	III	8			7.00	114.29
			27	6.75		
	IV	8			6.50	123.10
			25	6.25		
2001	I	4			5.88	68.03
			22	5.50		
	II	5			5.25	95.24
			20	5.00		
	III	5				
			–			
	IV	6				

Step 5: In this step, we calculate the modified mean for each quarter and take their sum. To obtain the modified mean, we arrange the percentages quarter-wise as shown in the table 2. Then we cancel out two extreme values for each quarter. That is, the highest and the lowest of the percentage values. Then the mean of the remaining values gives us the modified mean for that quarter respectively. In our example, after the cancellation of the highest and the lowest values, we are left with a single value. Therefore, this happens to be the mean value.

Table 2

Year	Quarter I	Quarter II	Quarter III	Quarter IV
1998	–	–	114.30	~~50.00~~
1999	~~133.30~~	~~72.73~~	~~118.34~~	68.49
2000	109.10	~~105.60~~	~~114.29~~	~~123.10~~
2001	~~68.03~~	95.24	–	–

Modified Means

Quarter I	:	109.10/1	=	109.10
Quarter II	:	95.24/1	=	95.24
Quarter III	:	114.30/1	=	114.30
Quarter IV	:	68.49/1	=	68.49
Total			=	387.13

The cancelling out of the highest and the lowest values reduces the effect of extreme cyclical and irregular variations and the averaging further smoothens the series for the same. The resultant values indicate the seasonal component.

Above, we have referred to the sum of modified means in each quarter. This sum should be ideally equal to 400 as each quarter has an index of 100. In our example, we observe that the sum of the modified means comes to 387.12. How it is adjusted constitutes our sixth step.

Step 6: In this step, we calculate the ratio of 400 and the sum of the modified means, that is 387.12. It comes out to be 1.0332. This value is referred to as adjusting factor. In order to distribute this deviation over all the means, we multiply each of the four modified means by the adjusting factor. After this correction, the sum of seasonal indices will add up to 400, giving an index of 100 for each quarter. Thus, the four indices in table 3 give us the components of the seasonal variation in the time series.

Table 3

Adjusting Constant = 400/387.13 = 1.0332

Quarter	Unadjusted Means x Adjusting Constant	=	Seasonal Index
I	109.10 1.0332	=	112.72
II	95.24 1.0332	=	98.40
III	114.29 1.0332	=	118.10
IV	68.49 1.0332	=	70.78
Total		=	400.00

Deseasonalizing a Time Series

The Ratio to Average Method allows us to identify the components of the seasonal variation in time series data, and the indices themselves help us to nullify the effects of seasonality on the time series. The use of indices to nullify the seasonal effects in the common parlance is referred to as Deseasonalizing the time series**.** Deseasonalizing a time series involves dividing the original data points with the relevant seasonal index expressed as a percentage.

In our example, the deseasonalization process is carried out as follows:

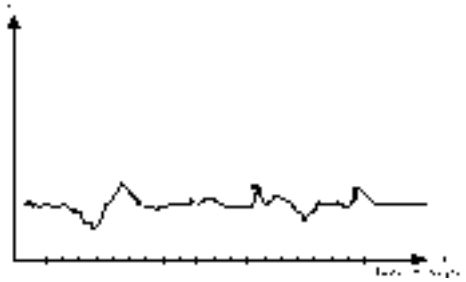
Year	Quarter	Actual Data	Seasonal Index/100	Depersonalized Data
(1)	(2)	(3)	(4)	(5) = (3)/(4)
1998	I	1	112.72/100	0.890
	II	2	98.40/100	2.033
	III	2	118.10/100	1.693
	IV	1	70.78/100	1.412

The removal of the seasonal component helps us to analyze the components of secular, cyclical and irregular variations. The secular trend then obtained can be utilized for projecting the trend into the future.

Irregular Variation

As the name suggests, the movement of the variable is random in nature without consistency and therefore, highly unpredictable. Since this type of irregularity exists for very short durations, the period under consideration will be of days, weeks and at the most of months. An appropriate example would be, a sudden spurt in the price of the share of a company rumors of a takeover. The irregular variation is shown in the figure below. Irregular variation is so inconsistent, that there is not a suitable mathematical model which would explain the phenomena under consideration.

Figure 5



The above figure exhibits the case of irregular variation.

COMPREHENSIVE ILLUSTRATION

Till now we have expressed the variations in terms of only one variation throughout. In the real world, this is seldom true and more often we find time series exhibiting more than one type of variation. In this part, we look at an example and examine it for three components of variation leaving out irregular variation.

Illustration 6

The Production Manager at Andhra Paper Mills Ltd., has the following data regarding the quantity of paper processed in thousand tonnes.

Year	Quarter I	Quarter II	Quarter III	Quarter IV
1998	3.1	5.1	5.6	3.6
1999	3.3	5.1	5.8	3.7
2000	3.4	5.3	6.0	3.8
2001	3.7	5.4	6.1	3.9

Given this data, we will

i. Deseasonalize it,

ii. Develop a trend line, and finally

iii. Find the component of the cyclical variation around the trend line, using Relative Cyclical Residual method.

i. **Table 4**

Year	Quarter	Quantity Processed	Four Quarter Moving Total	Four Quarter Moving Average	Centered Moving Avg. (CMA)	Percentage of Actual to CMA
	(1)	(2)	(3)	(4)	(5)	(6)
1998	I	3.1				
			–			
	II	5.1				
			17.4	4.35		
	III	5.6			4.38	127.85
			17.6	4.40		
	IV	3.6			4.40	81.82
			17.6	4.40		
1999	I	3.3			4.43	74.49
			17.8	4.45		
	II	5.1			4.47	114.10
			17.9	4.48		
	III	5.8			4.49	129.18
			18.0	4.50		
	IV	3.7			4.53	81.68
			18.2	4.55		
2000	I	3.4			4.58	74.24
			18.4	4.60		
	II	5.3			4.62	114.72
			18.5	4.63		
	III	6.0			4.67	128.48
			18.8	4.70		
	IV	3.8			4.72	80.51
			18.9	4.73		

2001	I	3.7			4.74	78.05
			19.0	4.75		
	II	5.4			4.77	113.21
			19.1	4.78		
	III	6.1				
			–			
	IV	3.9				

Table 5

Year	Quarter I	Quarter II	Quarter III	Quarter IV
1998	–	–	~~127.85~~	~~81.82~~
1999	74.49	114.10	~~129.18~~	81.68
2000	~~74.24~~	~~114.72~~	128.48	~~80.51~~
2001	~~78.05~~	~~113.21~~	–	–

Modified Means

Quarter I	:	74.49/1	=	74.49
Quarter II	:	114.10/1	=	114.10
Quarter III	:	128.48/1	=	128.48
Quarter IV	:	81.68/1	=	81.68
Total			=	398.75

Table 6

Adjusting Constant = 400/398.75 = 1.0031

Quarter	Unadjusted Means	x Adjusting Constant	=	Seasonal Index
I	74.49	1.0031	=	74.72
II	114.10	1.0031	=	114.45
III	128.48	1.0031	=	128.88
IV	81.68	1.0031	=	81.93
Total				399.98
			≃	400.00

Now, we Deseasonalize the given Time Series

Year	Quarter	Actual Data	Seasonal index/100	Deseasonalized Data
(1)	(2)	(3)	(4)	(5) = (3)/(4)
1998	I	3.1	74.72/100	4.149
	II	5.1	114.45/100	4.456
	III	5.6	128.88/100	4.345
	IV	3.6	81.93/100	4.394
1999	I	3.3	74.72/100	4.416
	II	5.1	114.45/100	4.456
	III	5.8	128.88/100	4.500
	IV	3.7	81.93/100	4.516

Year	Quarter	Actual Data	Seasonal index/100	Deseasonalized Data
2000	I	3.4	74.72/100	4.550
	II	5.3	114.45/100	4.631
	III	6.0	128.88/100	4.655
	IV	3.8	81.93/100	4.638
2001	I	3.7	74.72/100	4.952
	II	5.4	114.45/100	4.718
	III	6.1	128.88/100	4.733
	IV	3.9	81.93/100	4.760

Thus, the values given in column (5) are deseasonalized values. These values are used while fitting a secular trend and the cyclical variation.

ii. We now obtain a best fit for the secular trend. The working is shown below:

Year	Quarter	Deseasonalized Values (Y)	Coding the Time Variable	x = 2 . Col (4)	xY	x^2	Estimated Values $\hat{Y}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1998	I	4.149	–7(1/2)	–15	–62.35	225	4.2755
	II	4.456	–6(1/2)	–13	–57.928	169	4.3121
	III	4.345	–5(1/2)	–11	–47.795	121	4.3487
	IV	4.394	–4(1/2)	–9	–39.546	81	4.3853
1999	I	4.416	–3(1/2)	–7	–30.912	49	4.2190
	II	4.456	–2(1/2)	–5	–22.280	25	4.4585
	III	4.500	–1(1/2)	–3	–13.500	9	4.5951
	IV	4.516	–1/2	–1	–4.516	1	4.5317
2000	I	4.550	1/2	1	4.550	1	4.5683
	II	4.631	1(1/2)	3	13.893	9	4.6049
	III	4.655	2(1/2)	5	23.275	25	4.6415
	IV	4.638	3(1/2)	7	32.466	49	4.6781
2001	I	4.952	4(1/2)	9	44.568	81	4.7147
	II	4.718	5(1/2)	11	51.898	121	4.7513
	III	4.733	6(1/2)	13	61.529	169	4.7879
	IV	4.760	7(1/2)	15	71.400	225	4.8245
Total		72.869			24.867	1360	

$$b = \frac{\sum xY}{\sum x^2} = \frac{24.867}{1360} = 0.0183;$$

$$a = \overline{Y} = \frac{72.869}{16} = 4.55$$

Therefore, the estimating equation is given by

$$\hat{Y} = 4.55 + 0.0183x$$

iii. The calculation of cyclical variation by Relative Cyclical Residual Method is shown below:

Year	Quarter	Y	$\hat{Y}$	$\left(\frac{Y-\hat{Y}}{\hat{Y}}\right)100$
1998	I	4.149	4.2755	–2.96
	II	4.456	4.3121	3.34
	III	4.345	4.3487	–0.085
	IV	4.394	4.3853	0.198
1999	I	4.416	4.2190	–0.133
	II	4.456	4.4585	–0.056
	III	4.500	4.5951	0.109
	IV	4.516	4.5317	–0.346
2000	I	4.550	4.5683	–0.401
	II	4.631	4.6049	0.567
	III	4.655	4.6415	0.290
	IV	4.638	4.6781	–0.857
2001	I	4.952	4.7147	5.033
	II	4.718	4.7513	–0.701
	III	4.733	4.7879	–1.147
	IV	4.760	4.8245	–1.337

The components of cyclical variation are given in the last column of the above table.

TIME SERIES ANALYSIS IN FORECASTING

Forecasting is based on Time Series Analysis. Forecasting means "an art of making an estimate of future conditions on a systematic basis using available knowledge and information i.e., it is a rationally worked out estimate about future. Forecasting is done on specified assumptions and is always made with probability ranges."

From the previous discussion, it is clearly understood that the entire process of Time Series Analysis is to understand the behavior of a variable, so that a reasonably accurate estimate can be made. Hence, the approach should not be used mechanically. However, this can be a base to start with and amendments may have to be made depending on the context and views of the experts, if any in the process of "forecasting." Since the future is characterized by uncertainty, the need for forecasting has felt immensely, specially, in the areas of business where decisions have to be taken with regard to profits, sales, marketing strategies etc., by the management. Analysis of Time Series helps managers to make appropriate forecasts. Hence, whenever managers employ this technique, they should pay attention to the whole process and ascertain facts like authenticity of the historical data used, the variables which are expected to be erratic and hence change over the period during which the projections are made. Also changes in the internal process of the firm should be accounted for in the forecasting exercise.

ADDITIONAL ILLUSTRATIONS

Illustration 1

Cleanenviron Ltd., was able to sell the following number of devices during the past seven years

Years	1992	1993	1994	1995	1996	1997	1998
No. of devices sold '00	11	13	14	16	18	20	23

These devices reduce the amount of pollutants released by the tannery and other industries. With stricter enforcement of laws, the sales are expected to pick up. For this data, fit a

a. Linear estimating equations.

b. Curvilinear estimating equations.

c. Calculate the number of devices that will be sold in 1999 by using both the estimating equations. Which of the two estimates is more accurate?

Solution

a.

Years (X)	No. of Devices Sold (Y)	x = X – x	x.Y	x^2
1992	11	–3	–33	9
1993	13	–2	–26	4
1994	14	–1	–14	1
1995	16	0	0	0
1996	18	1	18	1
1997	20	2	40	4
1998	23	3	69	9
	N = 115		54	28

The mean of X $= \frac{13,965}{7} = 1995$

The slope of b $= \frac{\sum xY}{\sum x^2} = \frac{54}{28} = 1.93$

And a = $\bar{Y} = \frac{115}{7} = 16.43$

The regression equation describing the secular trend is given by

$\bar{Y} = 16.43 + 1.93x$

b. **Estimation of the Curvilinear Trend**

Years (X)	No. of Devices Sold (Y)	x	x^2	x^4	xY	x^2Y
1992	11	–3	9	81	–33	99
1993	13	–2	4	16	–26	52
1994	14	–1	1	1	–14	14
1995	16	0	0	0	0	0
1996	18	1	1	1	18	18
1997	20	2	4	16	40	80
1998	23	3	9	81	69	207
Total	115		28	196	54	470

The mean of X $= \frac{13,965}{7} = 1995$

$$\sum Y = an + c\sum x^2$$

We have $\sum x^2 Y = a\sum x^2 + c\sum x^4$ and $b = \frac{\sum xY}{\sum x^2}$

Substituting the values in these equations, we have

$115 = 7a + 28c$ … (1)

$470 = 28a + 196c$ … (2)

$b = \frac{54}{28} = 1.93$ … (3)

Multiplying (1) by 4, we have

$460 = 28a + 112c$ … (4)

We solve (4) and (2)

$460 = 28a + 112c$

$470 = 28a + 196c$

By subtractions, we have $10 = 84c$

Therefore, $c = (10/84) = 0.12$

Subsitituting the value of $c = 0.12$ in (1), we have

$115 = 7a + 28\ (0.12)$

$115 = 7a + 3.48$

$115 - 3.48 = 7a$

$7a = 111.52$

$a = 111.52 / 7 = 15.93$

On substituting the value of a, b and c in the equation of the parabola, we have

$\bar{Y} = 15.93 + 1.93x - 0.12x^2$

c. We substitute X = 1999 after translating it. It will be 1999 – 1995 = 4.

The number of devices that will be sold in this by linear equation is

$\bar{Y} = 16.43 + 1.93\ (4)$

$= 16.43 + 7.72$

$= 24.15$

That is in the current year, 2415 devices will be sold.

If we employ the curvilinear trend, the number of devices that sold are

$\bar{Y} = 15.93 + 1.93\ (4) - 0.12\ (4)^2$

$= 15.93 + 7.72 - 1.92$

$= 21.73$

That is in the current year, 2173 devices will be sold.

We are aware of the fact that the life cycle of any product does not extend for infinite period. Therefore, the sales forecast made based on the curvilinear trend may not be a prudent decision, although in this problem the estimate obtained from the curvilinear trend is more conservative as compared to the estimate obtained by the linear trend.

Illustration 2

The level of working capital needed by a small firm during the last six years us shown below:

Year	1993	1994	1995	1996	1997	1998
Working Capital In Rs.Lakhs	5.10	5.85	6.35	6.65	7.10	7.35

For this data, fit a linear trend and estimate the level of working capital required during the years 2000 and 2001 respectively.

Solution

Year(X)	Y	$x = X - \bar{X}$	X	XY	x^2
1993	5.10	–2.5	–5	–25.50	25
1994	5.85	–1.5	–3	–17.55	9
1995	6.35	–0.5	–1	–6.35	1
1996	6.65	0.5	1	6.65	1
1997	7.10	1.5	3	21.30	9
1998	7.35	2.5	5	36.75	25
11,973	38.40	0	0	15.3	70

$$\bar{X} = \frac{\sum x}{n} = \frac{11,973}{6} = 1995.5$$

$$\bar{Y} = \frac{\sum y}{n} = \frac{38.4}{6} = 6.4$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{15.3}{70} = 0.218571, \quad a = \bar{Y} = 6.4$$

The equation of linear trend would be therefore

$$\bar{Y} = 6.4 + 0.2186x$$

We now substitute the year 2000 after translating it in the estimating equation. That will be 2000 – 1995.5 = 4.5. This value multiplied by 2 will give us 9 half-year intervals. We have,

$$\bar{Y} = 6.4 + 0.2186x$$

$$\bar{Y} = 6.4 + 0.2186\ (9)$$

$$\bar{Y} = 8.367$$

That is, in the year 2000, the working capital requirement is expected to be Rs.8.367 lakh.

Now, we substiute the year 2001 after translating it in the estimating equation. That will be 2001 – 1995.5 = 5.5. this value multiplied by 2 will give us 11 half-year intervals. Therefore, we have

$$\bar{Y} = 6.4 + 0.2186x$$

$$\bar{Y} = 6.4 + 0.2186\ (11)$$

$$\bar{Y} = 8.805$$

That is, the working capital required in the year 2001 is expected to be Rs.8.805 lakh.

Illustration 3

For the above problem (No. 12.2) calculate the percent of trend and the relative cyclical residual.

Solution

The calculation of Percent of Trend and the Relative Cyclical Trend is shown bellow:

X	Y	$\bar{Y}$	% of Trend	Relative Cyclical Residual
1993	5.10	7.51486	67.86554	–32.1345
1994	5.85	7.67881	76.18368	–23.8163
1995	6.35	7.78811	81.53454	–18.4655
1996	6.65	7.85369	84.67357	–15.3264
1997	7.10	7.95206	89.28504	–10.7150
1998	7.35	8.00671	91.79800	–8.2020

Illustration 4

As a finance manager, you are responsible to raise funds from different sources so that the cost of capital is minimum possible. The first step you would take is to ascertain the amount of funds that you need in the current financial year. Towards this end, you collected the following information and obtained an equation for best linear fit as

Years	1992	1993	1994	1995	1996	1997	1998
Amount required (in lakh)	65	71	79	82	90	102.50	105

a. Calculate the Percent of Trend for this data.

b. Calculate the Relative Cyclical Residual trend.

c. In which year do you observe the largest fluctuation from the trend line? Was it same in both the cases?

Solution

The trend equation for this problem by best linear fit is

$$\bar{Y} = 84.93 + 6.93x$$

X	Y	$\bar{Y}$	% of Trend	Relative Cyclical Residual
1992	65	64.14	101.3408	1.340817
1993	71	71.07	99.90151	–0.09849
1994	79	78.00	101.2821	1.282051
1995	82	84.93	96.5501	–3.4499
1996	90	91.86	97.97581	–2.02482
1997	102.5	98.79	103.7554	3.755441
1998	105	105.72	99.31896	–0.68104

The maximum deviation by both the methods was found in the year 1997.

Illustration 5

Mr. Prasad, the Accounts Manger of Xcel Ltd., has complied the following data regarding the level of accounts receivable over a period of five years in Rs.(in Thousands).

Year	Quarter I	Quarter II	Quarter III	Quarter IV
1994	102	120	90	78
1995	110	126	95	83
1996	111	128	97	86
1997	115	135	103	91
1998	122	144	110	98

For this data, calculate the seasonal indices. After deseasonalizing the data, fit a linear trend and also calculate the cyclical components of variation by percent of trend method.

Solution

Year	Quarter	Amount Exported	Four Quarter Moving Average	Four Quarter Moving Average	Central Moving Average	% of Actual to CMA
1994	I	102				
			–			
	II	120				
			390	97.50		
	III	90			98.50	91.37
			398	99.50		
	IV	78			100.25	77.81
			404	101.00		
1995	I	110			104.13	105.64
			409	102.25		
	II	126			102.86	122.47
			414	103.50		
	III	95			103.63	91.67
			415	103.75		
	IV	83			104.00	79.80
			417	104.25		
1996	I	111			104.50	106.22
			419	104.75		
	II	128			105.13	121.75
			422	105.50		
	III	97			106.00	91.50
			426	106.50		
	IV	86			107.38	80.08
			433	108.25		

Year	Quarter	Amount Exported	Four Quarter Moving Average	Four Quarter Moving Average	Central Moving Average	% of Actual to CMA
1997	I	115			109.00	105.50
			439	109.75		
	II	135			110.38	122.30
			444	111.00		
	III	103			111.88	92.06
			451	112.75		
	IV	91			113.88	79.91
			460	115.00		
1998	I	122			115.88	105.28
			467	116.75		
	II	144			117.63	122.42
			474	118.50		
	III	110				
	IV	98				

Table 2

Year	Quarter I	Quarter II	Quarter III	Quarter IV
1994	–	–	~~91.37~~	~~77.81~~
1995	105.64	~~122.47~~	91.67	79.80
1996	~~106.22~~	121.75	91.50	~~80.08~~
1997	105.50	~~122.30~~	~~92.06~~	79.91
1998	~~105.28~~	122.42		–

Modified Means

Quarter I : 211.14 / 2 = 105.57

Quarter II : 244.17 / 2 = 122.09

Quarter III : 183.17 / 2 = 91.59

Quarter IV : 159.71 / 2 = 79.86

= 399.11

Table 3

Adjusting Constant = 400/399.11 = 1.0022

Quarter	Unadjusted Means x Adjusted Constant	Seasonal Index
I	105.57 x 1.0022	105.80
II	122.09 x 1.0022	122.36
III	91.59 x 1.0022	91.80
IV	79.86 x 1.0022	80.04
		400.00

Deseasonalizing the given Time Series

Year	Quarter	Amount Exported	Seasonal Index/100	Deseasonalized Data
1994	I	102	105.80/100	96.41
	II	120	122.36/100	98.07
	III	90	91.80/100	98.04
	IV	78	80.04/100	97.45
1995	I	110	105.80/100	103.97
	II	126	122.36/100	102.97
	III	95	91.80/100	103.48
	IV	83	80.04/100	103.69
1996	I	111	105.80/100	104.91
	II	128	122.36/100	104.61
	III	97	91.80/100	105.66
	IV	86	80.04/100	107.45
1997	I	115	105.80/100	108.68
	II	135	122.36/100	110.33
	III	103	91.80/100	112.20
	IV	91	80.04/100	113.69
1998	I	122	105.80/100	115.31
	II	144	122.36/100	117.69
	III	110	91.80/100	119.83
	IV	98	80.04/100	122.44

Thus, the values given in column (5) are deseasonalized values. These values are used in fitting a secular trend and the cyclical variations.

b. Obtaining a best fit for the secular trend. The working is shown below:

year	Quarter	Deseasonalized Data (Y)	Coding the Time	X	Xy	X^2
1994	I	96.41	–9 ½	–19	1831.8	361
	II	98.07	–8 ½	–17	–1667.2	289
	III	98.04	–7 ½	–15	–1470.6	225
	IV	97.45	–6 ½	–13	–1266.9	169
1995	I	103.97	–5 ½	–11	–1143.7	121
	II	102.97	–4 ½	–9	–926.73	81
	III	103.48	–3 ½	–7	–724.36	49
	IV	103.69	–2 ½	-5	–518.45	25
1996	I	104.91	–1 ½	–3	–314.73	9
	II	104.61	–0.5	–1	–104.61	1
	III	105.66	0.5	1	105.61	1
	IV	107.45	1 ½	3	322.35	9
1997	I	108.68	2 ½	5	543.45	25
	II	110.33	3 ½	7	772.31	49
	III	112.20	4 ½	9	1009.80	81
	IV	113.69	5 ½	11	1250.59	121
1998	I	115.31	6 ½	13	1499.03	169
	II	117.69	7 ½	15	1765.35	225
	III	119.83	8 ½	17	2037.11	269
	IV	122.44	9 ½	19	2326.36	361
		2146.89			1662.98	2660
Average		107.345				

$$a = \bar{Y} = \frac{2146.89}{20} = 107.345$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{1662.98}{2660} = 0.625180$$

Therefore, the linear trend is given by the equation

$\bar{Y} = 107.345 + 0.63x$

Year	Quarter	x	Deseasonalized y	$\bar{Y}$	% of Trend
1994	I	–19	96.41	95.375	101.0852
	II	–17	98.07	96.635	101.4840
	III	–15	98.04	97.895	100.1484
	IV	–13	97.45	99.155	98.2805
1995	I	–11	103.97	100.415	103.5403
	II	–9	102.97	101.675	101.2737
	III	–7	103.48	102.935	100.5295
	IV	–5	103.69	104.195	99.5153
1996	I	–3	104.91	105.455	99.4832
	II	–1	104.61	106.715	98.0275
	III	1	105.66	108.275	97.5849
	IV	3	107.45	109.235	98.3659
1997	I	5	108.68	110.495	98.3664
	II	7	110.33	111.755	98.7249
	III	9	112.20	113.015	99.2789
	IV	11	113.69	114.275	99.4881
1998	I	13	115.31	115.535	99.8053
	II	15	117.69	116.795	100.7663
	III	17	119.83	118.055	101.5035
	IV	19	122.44	119.315	102.6191

Illustration 6

The exchange rates (Rs./Dollar) for the last two years are given below:

Month	Exchange rate (Rs./Dollar)	
	1997-98	1998-99
April	35.80	39.63
May	35.79	40.45
June	35.80	42.18
July	35.72	42.44
August	35.88	42.69
September	36.40	42.46
October	36.80	42.28
November	37.18	42.33
December	39.09	42.52
January	39.21	42.46
February	38.83	
March	39.49	

Calculate the quarterly moving average of exchange rates from June,1998 to December, 1998.

Solution

The data required for calculating the quarterly moving average from June,1998 to December 1998, along with the quarterly moving average of exchange rate is given below:

Month	Exchange Rate	Quarterly Moving Average
April	39.63	–
May	40.45	–
June	42.18	40.75
July	42.44	41.69
August	42.69	42.44
September	42.46	42.53
October	42.28	42.47
November	42.33	42.35
December	42.52	42.37

Illustration 7

The following data pertains to the estimated demand and actual demand for a product for the last eight periods.

Period	Estimated Demand ('000 units)	Actual Demand ('000 units)
1	714	750
2	546	526
3	568	652
4	739	667
5	577	682
6	662	859
7	788	1022
8	1097	981

You are required to identify the cyclical variations in demand using percent of trend measure.

Solution

Period	Estimated Demand ('000 Units) $\overline{Y}$	Actual Demand ('000 Units) Y	% of Trend x 100
1	714	750	105.04
2	546	526	96.34
3	568	652	114.79
4	739	667	90.26
5	577	682	118.20
6	662	859	129.76
7	788	1022	129.70
8	1097	981	89.43

SUMMARY

- Time series analysis is used to identify and determine the pattern of changes in the data collected over regular intervals of time. The identified patterns are projected into future to get an estimate of the variable under consideration.
- The variations observed in the time series can be broadly classified as (i) the secular trend, (ii) the cyclical fluctuation, (iii) the seasonal variation, and (iv) the irregular variation.
- Secular trend observes the long-term behavior of the variable. Graphically, the secular trend is shown as a straight line with an upward slope, with the actual time series represented as a curve moving towards and away from the trend line. A best fit trend line can be obtained by using the least squares method. Secular trend is useful in examining whether a policy implemented has yielded the necessary results or not and its future impact in estimating the variable under study; in studying any other components present in the series etc.
- By cyclical variations, we refer to the long-term movement of the variable about the trend line. Generally, the behavior of the variable is considered to be cyclic only if the movements recur after a period of more than one year. The Residual method is used to isolate the components of cyclical variation in a time series, and this method can be bifurcated into two measures: percent of trend and relative cyclical residual measures.
- Under seasonal variation, we observe the similar pattern that the variable under consideration shows during certain months of the successive years. Usually, the time period over which this variation is considered can consist of days, weeks, months and at the most one year. The ratio to moving averages method is used to identify the components of seasonal variation in a time series. The use of indices to nullify the seasonal effects is referred to as deseasonalizing the time series.
- Irregular variation is the movement of the variable in a random fashion without consistency, which makes it highly unpredictable. Since this type of irregularity exists for very short durations, the period under consideration will be of days, weeks and at the most of months.

Chapter XIII

Probability

After reading this chapter, you will be conversant with:

- The Concept of Probability
- Approaches to Probability
- Probability Rules
- Bayes' Theorem
- Additional Illustrations

Introduction

Probability has its origin from gambling theory/games of chance, which is based on the concept of 'chance' such as tossing the coin, playing the cards etc. 'Chance' is the essence of probability.

The first book on Games of Chance was written by Jerame Cardon, an Italian mathematician. He has given a number of rules for minimizing the risk of gambling and protecting the person against cheating. Probability as a quantitative measure was first attempt by an Italian mathematician, Gallieo. However, systematic and scientific foundation was given by Blaise Pascal and Pierre de Fermat, two French mathematicians, while solving an incomplete gambling problem for sharing the stake. Another stalwart was Swiss Mathematician James Bernoulli made extensive study of the subject and his 'Treatise on Probability' was published in 1713 is a major contribution to the theory of probability. A. De-Moivre also contributed a lot to the subject and published his work *The Doctrine of Chance* in 1718. The concept of *Inverse Probability* was introduced by Thomas Bayes in (1702-61). Theorie Analytique des Probability (Theory of Analytical Probability) an monumental work was published in 1812 by French Mathematician Pierre-Simon de Laplace after conducting an extensive research which resulted in the *Classical theory of Probability*. Then, the empirical approach to probability was introduced by R. A. Fisher and Von Mises through the notion of sample space. Russian mathematician contribution to the modern theory of probability is very great. The theory of probability was axiomized by Kologorov. He introduced probability as a set function in his book '*Foundation of Probability'* published in 1933.

The applications of probability in business have attained high significance in the recent past. With the growing emphasis on the management's role in the success of the company, the role of probability theory has become prominent. The utility of probability in business and economics can be seen in making future predictions. Uncertainty plays an important role in business and probability is a concept that measures the degree of uncertainty and that of certainty also as a corollary. The present chapter deals with the concept and the sub-concepts involved in it in detail.

THE CONCEPT OF PROBABILITY

A student is considering whether she should enroll in an educational program. Among other things, she would like to know how difficult the program is. She obtains the following marks distribution of students who appeared for the recent final exam.

Relative Frequency Distribution

Marks %	No. of Students	% of Students
0-25	45	8
25-50	280	50
50-75	205	37
75-100	30	5
	560	100

Assuming the next exam is equally tough and there is same proportion of dull and bright students, she can conclude that the percentage of students in the four classes of marks will again be

Marks %	% of Students
0-25	8
25-50	50
50-75	37
75-100	5
	100

The first distribution is related to past data and is a frequency distribution. The second distribution has the same numbers and is a copy of the first distribution. However, this distribution relates to the future. Such a distribution is called a **probability distribution**. Note the similarity of this distribution with that of the relative frequency distribution.

Hence, by inspecting the probability distribution, we can say that:

8% of the students who are appearing for the exam will score 0-25% marks, 50% will score 25-50% marks, 37% will score 50-75% marks and the balance 5% will score between 75-100% marks.

If our student considers herself to be among the top 5% of the students, she can conclude that she will score between 75 to 100%. If she considers herself to be in the top 42% of the students, she can conclude that she will score 50-100% marks and so on. However, if she has no idea of her ability in comparison to the other students, she can conclude that:

She has 8% chance of scoring 0-25% marks, a 50% chance of scoring 25-50% marks, a 37% chance of scoring 50-75% marks and a 5% chance of scoring 75-100% marks. This "chance" is called **probability** in statistical language. It is commonly used in day-to-day conversation. For example, the statements like "probably it may rain today", or "the chances of India winning the cricket match are equally good", etc., indicate uncertainty about happening of some events. So, in layman's words, probability is nothing but the uncertainty about the happening of an event. However, in Statistic and Mathematics, we present a condition under which some sensible numerical statements are made about the uncertainty and certain methods are applied for calculating the value of probabilities and their expectations.

Today, probability has become one of the tools of statistics. The subject has developed to a great extent and there is no such discipline where the theory of probability is not used. It is an essential tool of statistical inferences and forms the basis of Decision theory. In fact, statistics and probability are interrelated to each other.

Probability theory is used to analyze data for decision making.

Probability theory, as discussed earlier, provides solutions to the social, economic, political and business problems. The insurance industry uses probability theory to calculate premium rates; a stock analyst/investor, based on the probability estimates of economic scenarios, estimates the returns of the stocks; a project manager applies probability theory in decision making.

Terminology in Probability

Experiment is an operation that produces outcomes which can be observed. In other words, experiment is an act, which can be repeated under some essential homogenous conditions. An experiment is called a *Random Experiment* if, when conducted repeatedly under a given condition gives result, which depends on chance. The result of random experiment is called Outcome, which is not unique but may be one of various possible outcomes. In other words, random experiments are those experiments whose possible outcomes are known in advance and none of the outcomes are predicted with certainty.

Trial is performing a random experiment.

Outcome/Event is the result of an experiment. In other words, the outcome or the combinations of outcomes are called as Events. An event is a random event when it may or may not occur while performing a random experiment.

The following table depicts the various experiments conducted and their related outcomes:

Table 1: Various Experiments and their Relative Outcomes

	Experiment	Outcome
1.	Inspecting a light bulb	Defective or non-defective
2.	Examining a student's academic record	Cumulative grade point average
3.	Examining a medical record	Diagnosis of a disease
4.	Purchasing shares of a common stock	Increase, decrease or no change in price per share
5.	Rolling dice	Numbers appearing on the upturned faces
6.	Running a mile	Time elapsed
7.	Following a diet	Weight loss or gain
8.	Running for public office	Win or lose
9.	Purchasing a new automobile	Performance acceptable or unacceptable

The following are the different types of events:

- Events are said to be **mutually exclusive** if they cannot occur at the same time when an experiment is performed. Two or more events are said to be mutually exclusive, when both cannot happen simultaneously in a single trial i.e., happening of one event excludes the happening of another or all the events. Thus, if two events are mutually exclusive, the acceptance of one precludes the acceptance of another. For example, if a coin is tossed, the event either head or tail will come, as both cannot be up at the same time. Such events are known as Mutually Exclusive Events.
- **Independent and Dependent Events:** Two events, A and B, are **independent events** if the occurrence of event A is in no way related to the occurrence or non-occurrence of event B. Likewise, for independent events, the occurrence of event B is in no way related to the occurrence of event A. Two events A and B are **dependent events** if the occurrence of one event say A is related to the occurrence of another event, say B.
- **Equally Likely Events** – When the outcome is not expected to occur in preference to another such events are called as Equally Likely Events. In other words, one event does not occur more often than the other event. For example, in case of tossing of a coin or throwing of a die, all the outcomes (Head or Tail, or 1, 2, 3, 4, 5 and 6) are expected to occur for the same number of times in the long run.
- The happening or non-happening of a single event is a case of **simple event** and **compound event** is the result of happening or non-happening of two or more events. Drawing a blue ball form a bag containing 5 red balls and 5 blue balls is an example for simple event and for compound event, the example is as follows:

  If a bag contains 10 white balls and 6 red balls and if two successive draws of 3 balls are made, we shall be finding out the probability of getting 3 white balls in the first draw and 3 red balls in the second draw.

- **Collectively exhaustive** list is a list of all possible events of a random experiment. For example, if a single coin is tossed, the exhaustive number of cases is two i.e., we may get either head or tail. Similarly, if a die is thrown, the possible outcomes are 1, 2, 3, 4, 5 and 6 i.e., the exhaustive number of events is 6.
- If there are two events in an experiment, then one event is called the complimentary event of the other event, provided they are mutually exclusive and exhaustive.

  Sample Space is the totality of all possible outcomes of an experiment. The sample space is denoted by S. Each element in a sample space is called sample point or event point and is defined as "a non-empty sub-set of the sample space, i.e., out of all the possible outcomes of a random experiment, some outcomes satisfy a specified description, are called as Event."

  Elements of sample space/point possess the following properties:

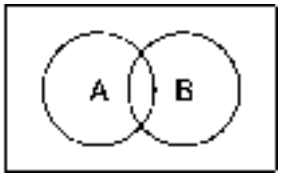
  i. Each sample point is an outcome of a random experiment.

  ii. If the experiment is repeated, then the resulting outcome corresponds to one and only one of the sample space.

  In the given table, if the experiment relates to inspecting a light bulb, the only possible outcome is that it is either defective or not defective. Hence, the sample space has two members, defective or not defective.
- **The Venn Diagram:** A Venn diagram is a pictorial representation of the sample space of an experiment. It is usually drawn as a rectangular figure representing the sample space and it contains circles or other shapes representing events in the sample space.

Figure 1

Venn diagram representing outcomes of tossing a coin viz., heads and tails.



Venn diagram representing outcomes of selecting a manager

A = candidate has over 3 years experience

B = candidate has post graduate qualification.

Example 1

A gambler places a bet on numbers 14 through 25. There are 12 equally likely winning outcomes. The roulette wheel (a gambling instrument which can display any one of the 38 equally likely numbers as the winning number) contains 38 equally likely outcomes.

The probability of the wheel stopping on a number from 14 through 25, say event A = P(A) = 12/38 = 0.316.

The probability of losing, i.e., the wheel stopping on numbers other than 14 through 25 (say event B) is the probability of the complement of A occurring. The complement of an event A is defined as A', where A' represents the non-occurrence of event A. So, the probability of A' (B) = 26/38 = 0.684.

$P(A')$ = P(B) = 1 – P(A) because A and A' are the only possible events and they are mutually exclusive events of the sample of 38 equally likely outcomes. Thus, P(A) + P (A') = 1 and P(A and A') is 0.

Definition of Probability

The word *probability* does not have a consistent direct definition. Actually, there are two broad categories of probability interpretations: Frequentists talk about probabilities only when dealing with well defined *random experiments*. The relative frequency of occurrence of an experiment's outcome, when repeating the experiment, is a measure of the probability of that random event. Bayesians, however, assign probabilities to *any statement whatsoever*, even when no random process is involved, as a way to represent its subjective plausibility.

Probability is defined as a chance of occurrence of an event. In other words, it is an expression of likelihood. Probability ranges between 0 and 1. A zero (0) probability is assigned to an event which cannot occur and for an event which is certain to occur, probability 1 is assigned.

APPROACHES TO PROBABILITY

There are four different schools of thought/approaches on the concept of probability. They are:

- Classical or a Priori Probability
- Relative Frequency Theory of Probability
- Subjective Approach to Probability
- Axiomatic Approach to Probability.

Classical Probability

According to the Classical approach, probability is the ratio of the number of equally likely possible outcomes favorable for an event to the total number of possible outcomes. If there are m possible outcomes that favor the occurrence of event A and there are n total possible outcomes, then the probability of the occurrence of event A is the ratio of m to n (m/n). The possible outcomes favorable for an event and total number of outcomes must be known without performing experiments.

Consider the experiment of tossing a single coin. Two outcomes are possible, viz. obtaining a head or obtaining a tail. The probability that it is a tail is 1/2, i.e., 0.5. This probability is determined without an experiment based on the principle that each of the possible outcomes must be equally likely. In reality it may not be that, for every two tosses there will be one tail. However, if the number of tosses is increased, the actual results will approximate more closely to the one-in-two pattern. Thus, in 2000 tosses, there may be 998 tails. If the probability of a tail being tossed is one in two, it does not follow that, and in order to maintain the probability ratio, the next toss will produce a head.

Relative Frequency Definition or Empirical Approach

This type of probability requires us to make some qualifications. We define probability of event A, occurring as the proportion number of times A occurs, if we repeat the experiment several times under the same or similar conditions.

According to Von Mises, "If the experiment is repeated a large number of times under essentially identical conditions, the limiting value of the ratio of the number of times the event A happens to the total number of trials of the experiment as the number of trials increases indefinitely is called the probability of happening of the event A."

Symbolically, Let P(A) denote the probability of the occurrence of A and m be the number of times in which an event A occurs in a series of n trials. Then,

$$P(A) = \lim_{n\to\infty} \frac{m}{n}$$ provided the limit is finite and unique.

Example 2

Consider the following data regarding distribution of salaries in a finance company for February, 2002.

Salaries (Rs.)	Frequency (m)	Relative Frequency (%)
2,000-5,000	2	4%
5,000-8,000	11	22%
8,000-11,000	18	36%
11,000-14,000	10	20%
14,000-17,000	7	14%
17,000-20,000	2	4%
	Total = n = 50	100%

For the subsequent month, the salaries are likely to have the same distributions unless employees leave or have their salaries raised, or new people join. Hence, we have the following probabilities obtained from the above relative frequencies.

Salaries (Rs.)	Probability
2,000-5,000	4%
5,000-8,000	22%
8,000-11,000	36%
11,000-14,000	20%
14,000-17,000	14%
17,000-20,000	4%
	100%

These probabilities give the chance that an employee chosen at random will be in a particular salary class. For example, the probability of an employee's salary being Rs.5,000-Rs.8,000 is 22%.

Subjective Probability

Probability may be determined by a personal statement of how likely an outcome is in a single trial or repetition of the same experiment.

Since subjective probabilities are based on personal judgment, they are peculiar to the individual making the decision. The probability statement depends upon the individual's experience and his familiarity with the facts of the case. Two decision makers with the same amount of information would make different subjective probability estimates of the occurrence of a particular event.

Example 3

An expert analyst of share prices may give his judgment that the price of ACC shares has a 20% probability of increase i.e., by Rs.500 or more in the next two months, a 60% probability of increase i.e., by less than Rs.500 in the next two months and a 20% probability of remaining unchanged or registering a slight fall.

Axiomatic Approach to Probability

The modern theory of probability is based on the axiomatic approach. The axiomatic approach to probability was introduced by A. N. Kolmogorov, a Russian Mathematician, in the year 1930. Kolmogorov axiomised the theory of probability and his book '*Foundation of Probability'* in 1933, introduced probability as a set of functions. In axiomatic approach, some concepts are laid down and certain properties or postulates known as axioms are defined on which probability

calculations are based. The entire theory is developed by logical deduction on the basis of postulates. This approach includes both the classical and empirical approach of probability but at the same time is free from their drawbacks.

Given the sample space of a random experiment, the probability of occurrence of an event is defined as a set of functions satisfying the following three axioms:

1. Probability of event ranges between 0 and 1. If the event does not occur, its probability is 0 and if the event is bound to occur, its probability is 1.
2. The probability of the entire sample space = 1 i.e., P(S) = 1.
3. If two events, A and B, are mutually exclusive events, then the probability of occurrence of either A or B is denoted by P (A $\cup$ B).It is given by

   P (A $\cup$ B) = P(A) + P(B).

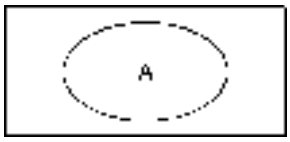
AXIOMS

A probability is a number assigned to the occurrence of an event in a sample space. Probability measures must satisfy three rules. If A is an event with probability denoted by P(A), then the following rules hold:

Axiom 1

The probability of the entire sample space S is 1, i.e., P(S) = 1.

Figure 2



Area of sample space rectangle = 1. An event A is represented within the rectangle. Minimum possible area of A is 0 and maximum possible area is 1.

Axiom 2

The probability of the event A must be greater than or equal to 0 and less than or equal to 1 or 100%, i.e., $0 \leq P(A) \leq 1$. This rule says that probabilities cannot be negative and as the probability of the sample space is 1, the probability of an event contained in the sample space should be less than or equal to 1.

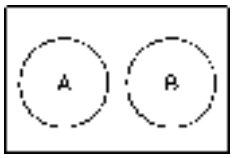
Axiom 3

If A and B are mutually exclusive events, then the probability of (A or B) is equal to the sum of the probabilities of A and B.

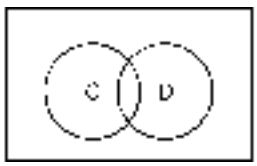
P(A or B) = P (A) + P (B) because P (A and B) = 0 as A and B are mutually exclusive.

Mutually exclusive events are those which do not overlap when represented in Venn diagrams.

Figure 3



A and B are Mutually Exclusive Events



C and D are not Mutually Exclusive Events

Two events, A and B, are mutually exclusive if the occurrence of one implies the non-occurrence of the other. Hence, obtaining a head and obtaining a tail on tossing a coin are mutually exclusive events.

PROBABILITY RULES

The Addition Rule

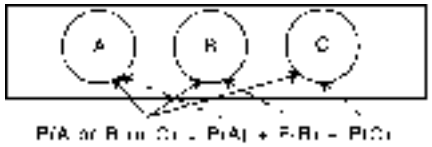
It is applicable to both mutually exclusive events and non-mutually exclusive events, and is based on Axiom 3.

MUTUALLY EXCLUSIVE EVENTS

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

This can be represented with the Venn diagram as follows:

Figure 4



Example 4

Suppose, A = getting 1 on throw of the die

B = getting 2 on throw of the die

C = getting 3 on throw of the die

As there are six possible equally likely outcomes on throwing the die,

$$P(A \text{ or } B \text{ or } C) = \frac{3}{6} = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = P(A) + P(B) + P(C).$$

NON-MUTUALLY EXCLUSIVE EVENTS

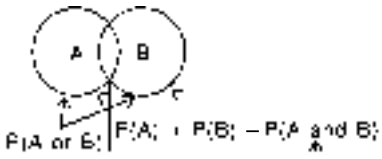
If two events are not mutually exclusive, the probability of one of them occurring is the sum of the marginal probabilities of the events minus the joint probability of the occurrence of the events.

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

Where, A and B are not mutually exclusive events.

A venn diagram illustrating the above is given below.

Figure 5



Example 5

Warwick Systems Company markets personal computers. Some computers have two disk drives (A) and some have one disk drive (B). Another feature of these machines is the capacity in terms of K (kilo) bytes – that is, whether they have 256K or 128K capacity. We will represent 256K by C and 128K by D. Presently, the firm's finished goods inventory consists of 300 machines equipped with varying features (see Table). At any time, Warwick Systems Company may receive an order for a machine or machines with specific features. If Warwick has a sufficient number of machines to satisfy its customers, the customers will continue to order machines from Warwick. However, if Warwick cannot satisfy its customers' needs, they will probably order machines elsewhere. Hence, the

management of Warwick wishes to know the likelihood that its inventories contain machines with desirable features.

	Two Disk Drives (A)			One Disk Drive (B)			Total
With 256K capacity (C)		100			50		150
With 128K capacity (D)		100			50		150
Total		200			100		300

In the above example, the sample space S is the set of all machines in inventory. What is the probability of a random selection of a two-disk drive machine from inventory, or P(A)?

$$P(A) = \frac{\text{Number of ways in which A occurs}}{\text{Number of outcomes in the sample space}} = \frac{200}{300} = \frac{2}{3}$$

The probability of randomly selecting a machine with 256K capacity is

$$P(C) = \frac{150}{300} = 0.5$$

Each of the above probabilities is designated as a **marginal**, or **unconditional probability**. Events A and C are not mutually exclusive since a machine may have both characteristics. The probability of a machine having two disk drives or having 256K capacity involves the addition rule with a twist. Since A and C are not mutually exclusive events, we must apply the counting rule. Hence, the probability of A or C is

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

$$= \frac{200}{300} + \frac{150}{300} - \frac{100}{300} = \frac{5}{6}$$

Event (A or C), then includes all elements except the 50 elements of B that are elements of neither A nor C.

The probability of a machine having features B and D is

$$P(B \text{ and } D) = \frac{50}{300} = \frac{1}{6}$$

The probability of the complement of (A or C) is P(B and D). These two events account for all 300 computers. The equation for P(A or C) sums the elements from A and C and subtracts elements that are in both A and C.

General Rule

A general rule is to subtract the probabilities with an even number of components inside the parentheses and add those with an odd number of components (one or three) to arrive at the probability of events.

Example 6

Consider a bag containing 4 white and 5 black balls. A man draws 3 balls at random; let us see what is the probability that all three balls are black?

The total number of ways in which 3 balls can be drawn is ($9c_3$) and the number of ways of drawing 3 black balls is ($5c_3$); therefore, the probability of drawing 3 black balls is given by

$$P(\text{All three are black}) = \frac{\text{Favorable Events}}{\text{Total Events}}$$

$$= \frac{5c_3}{9c_3} = \frac{\frac{5!}{2! \times 3!}}{\frac{9!}{6! \times 3!}} = \frac{10}{84} \text{ or } \frac{5}{42}$$

Example 7

Consider a bag containing 5 white and 7 black balls; if two balls are drawn, what is the probability that one is white and the other is black?

P (One is white and the other, black)

$$= \frac{\text{Favorable Events}}{\text{Total Events}} = \frac{5c_1 \times 7c_1}{12c_2} = \frac{\frac{5!}{4!\times 1!} \times \frac{7!}{6!\times 1!}}{\frac{12!}{10!\times 2!}} = \frac{5\times 7}{66} = \frac{35}{66}$$

Multiplication Rule

It is applicable for Independent and dependent events. Before actually discussing the multiplication rule, let us discuss Unconditional and Conditional probabilities.

MARGINAL PROBABILITY/UNCONDITIONAL PROBABILITY

Probability of event A happening, denoted by P(A), is called single probability, marginal or unconditional probability.

Marginal or Unconditional Probability is defined as the ratio of number of possible outcomes favorable to the event A (**favorable events**) to the total number of possible outcomes.

$$P(A) = \frac{\text{Number of possible outcomes favoring A}}{\text{Total number of possible outcomes}}$$

The definition assumes that the elements of the sample space have an equally likely chance of occurring.

CONDITIONAL PROBABILITY

If the probability of an event is subject to a restriction on the sample space, the probability is said to be conditional. Let us discuss this with the help of an example. Let there be two events viz., A and B. The probability of the happening of an event B when it is known that event A has already happened is called the conditional probability of B and is denoted by P(B/A).

UNCONDITIONAL PROBABILITY

Independence of Events

The joint probability of two outcomes that are independent is equal to the probability of the first outcome multiplied by the probability of the second outcome (the second outcome may be occurring simultaneously or sometime in the future), i.e., joint probability of two independent events is equal to the product of their marginal probabilities.

P(A and B) = P(A) . P(B)

Where, P(A) and P(B) are independent events.

Example 8

If the probability that A will be alive for more than 20 years is 0.7 and the probability that B will be alive for more than 20 years is 0.5, then the probability that they will both be alive for more than 20 years is 0.7 x 0.5 = 0.35.

Remark

We may extend the multiplication rule for independent events to three or more events by the following formula:

P(A and B and C….) = P(A) P(B) P(C) ….

Example 9

Consider a true-false test with 3 questions. Assuming that there is no idea of the true answers to the questions and each is guessed, we have a 50% chance of being correct. That is, the probability of success is 0.5 and probability of failure is 0.5 on each question.

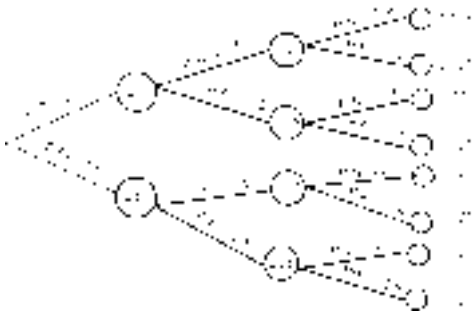
$P(S_1) = 0.5$ [S_1 = Successful in question 1]

$P(F_1) = 0.5$ [F_1 = Failing in question 1]

$P(S_1\ S_2) = P(S_1)\ P(S_2) = 0.5 \times 0.5 = 0.25$

We can illustrate this with the help of a tree diagram.

Figure 6



We multiply along the branches of the trees to get the joint probabilities.

For example, $P(F_1\ S_2\ F_3) = 0.5 \times 0.5 \times 0.5 = 0.125$.

Dependence of Events

The joint probability of two events A and B which are dependent is equal to the probability of A multiplied by the probability of B given that A has occurred.

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

or $P(B \text{ and } A) = P(B)\ .\ P(A|B)$

This formula is derived from the formula of conditional probability of dependent events.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

$\Rightarrow P(A \text{ and } B) = P(B|A).P(A).$

Example 10

A study of an insurance company shows that the probability of an employee being absent on any given day P(A) is 0.1. Given that the employee is absent, the probability of that employee being absent a second day in succession P(B|A) is 0.4. Events A and B are dependent events because B cannot occur unless event A has occurred. The probability of the employee being absent on two successive days

$P(A \text{ and } B) = P(A)\ .\ P(B|A)$

$= (0.1)\ (0.4) = 0.04$

Thus, the probability of the employee being absent on two successive days is 0.04 or 4% of the time.

Joint probability of several dependent events is equal to the product of the probabilities of occurrence of the preceding outcomes in the sequence.

$P(A \text{ and } B \text{ and } C...) = P(A)\ P(B|A)\ P(C|A \text{ and } B)$ ....

Example 11

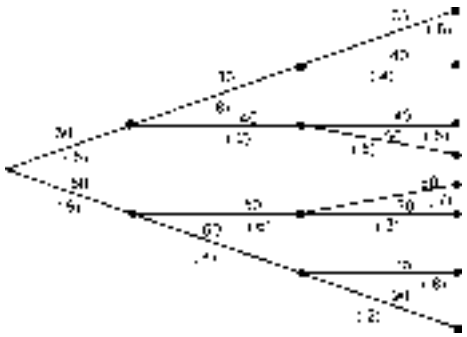
Let us consider a project which involves an outlay of Rs.1,00,000. The cash inflows expected to be generated by the project are shown in the figure below. From the figure we find that there are eight possible cash flow streams. The first cash flow stream consists of Rs.30,000 in year 1, Rs.30,000 in year 2 and Rs.35,000 in year 3, the second cash flow stream consists of Rs.30,000 in year 1, Rs.30,000 in year 2 and Rs.40,000 in year 3. So on and so forth. The probabilities associated with these cash flow streams are given in the parentheses. It may be noted that the probability with which a cash flow stream occurs is simply the joint probability of the individual elements in that cash flow stream.

Solution

The probability of the first cash flow stream, for example, is:

P(Rs.30,000 in year 1) x P (Rs.30,000 in year 2| Given Rs.30,000 in year 1) x P(Rs.35,000 in year 3| Given Rs.30,000 in year 1 and Rs.30,000 in year 2) = (0.5) (0.8) (0.6) = 0.24

Year 1		Year 2		Year 3			
Net Cash Flow	Initial Probability P(1)	Net Cash Flow	Conditional Probability P(2|1)	Net Cash Flow	Conditional Probability P(3|2, 1)	Cash Flow Stream	Joint Probability P(1,2,3)
				35,000	0.6	1	0.24
		30,000	0.8	40,000	0.4	2	0.16
30,000	0.5	40,000	0.2	45,000	0.5	3	0.05
				50,000	0.5	4	0.05
				60,000	0.7	5	0.21
50,000	0.5	50,000	0.6	70,000	0.3	6	0.09
		60,000	0.4	75,000	0.8	7	0.16
				90,000	0.2	8	0.04



Marginal probability in case of independent events is just the addition of the probabilities of all the events in which the simple event occurs.

Example 12

In the cash flow streams problem, we can find the probability of cash inflow of Rs.30,000 in year 1 given the joint probabilities of cash flow streams involving the cash inflow of Rs.30,000 in year 1, i.e., of years 1, 2, 3 and 4.

P(Rs.30,000, Rs.30,000, Rs.35,000 in years 1, 2, and 3) = 0.24

P(Rs.30,000, Rs.30,000, Rs.40,000 in years 1, 2 and 3) = 0.16

P(Rs.30,000, Rs.40,000, Rs.45,000 in years 1, 2 and 3) = 0.05

P(Rs.30,000, Rs.40,000, Rs.50,000 in years 1, 2 and 3) = 0.05

The probability of the cash inflow of Rs.30,000 in year 1

= 0.24 + 0.16 + 0.05 + 0.05 = 0.5

Example 13

Suppose that a sample of size 2 is chosen from a population of 6 elementary units. The sampling is performed without replacement. Thus, an element of the population can only be selected once in a sample.

Each possible sample of size 2 has the same chance of being selected.

Let the elementary units of the population be denoted by A, B, C, D, E, and F. Then, the possible samples of size 2 are:

AB	AE	BD	CD	DE
AC	AF	BE	CE	DF
AD	BC	BF	CF	EF

Each of these 15 equally likely samples of size 2 has a probability of being selected.

Consider the sample denoted by CE. Units C and E can be selected in any order. We consider the order CE and EC as separate events. The probability of selecting C and then E is

P(C and E) = P(C) P(E|C) = (1/6) (1/5) = 1/30

Likewise, the probability of selecting E and then C is

P(E and C) = P(E) P(C|E) = (1/6) (1/5) = 1/30

These two joint events are mutually exclusive, and the probability of one or the other occurring is

$$P[(C \text{ and } E) \text{ or } (E \text{ and } C)] = \frac{1}{30}+\frac{1}{30}=\frac{2}{30}=\frac{1}{15}$$

This value is the probability of C and E occurring in any order. In the same manner, we may show that the probability of any such sample is 1/15.

Remark

The above probability calculation is a combination of both the multiplication rule and the rule for mutually exclusive events. We use the multiplication rule to find the joint probability of particular outcomes or events. In turn, we use the rule of addition for mutually exclusive events to find the probability of one or the other of the two joint events occurring. Thus, both rules may be used in the same problem to answer different questions.

CONDITIONAL PROBABILITY

Independence of Events

Conditional probability, as said earlier, is the probability of the occurrence of an event, say A, subject to the occurrence of a previous event, say B. We define the conditional probability of event A, given that B has occurred, in case of A and B being independent events, as the probability of event A.

P(A|B) = P(A)

It is so because independent events are those whose probabilities are in no way affected by the occurrence of each other.

Example 14

Let us take the same true-false test. As the success answers are independent of each other we can say that the probability of success of the second answer given that the first answer is a success is simply the probability of the success of the second answer, i.e.,

$$P(S_2|S_1) = P(S_2) = 0.5$$

Dependence of Events

We can define the conditional probability of event A, given that event B occurred when both A and B are dependent events, as the ratio of the number of elements common in both A and B to the number of elements in B.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Example 15

The data regarding the salary and wage of workers is given below and the conditional probability is also calculated.

Membership Status	Non-agricultural Industries (B_1)	Agricultural Industries (B_2)	Total
Members of labor organizations (A_1)	20,044	51	20,095
Non-members represented by labor organizations (A_2)	2,394	4	2,398
Non-members not represented by labor organizations (A_3)	63,586	1,400	64,986
Total	86,024	1,455	87,479

Event A_1 denotes members of labor organizations. The probability of an employed worker being a member of a labor organization (event A_1) is

$$P(A_1) = \frac{20,095}{87,479} = 0.230$$

The probability of a worker being employed in a non-agricultural industry (event B_1) is

$$P(B_1) = \frac{86,024}{87,479} = 0.983$$

Now, we wish to determine the probability that a worker is a member of a labor organization given that the worker is employed in a non-agricultural industry. So, we must calculate the conditional probability of event A_1 occurring given that event B_1 has occurred. The formula for the conditional probability is

$$P(A_1|B_1) = \frac{P(A_1 \text{ and } B_1)}{P(B_1)}$$

The probability of a worker being both a member of a labor organization and employed in a non-agricultural industry is:

$$P(A_1 \text{ and } B_1) = \frac{20,044}{87,479} = 0.229$$

The conditional probability is then computed as

$$P(A_1|B_1) = \frac{20,044/87,479}{86,024/87,479} = \frac{20,044}{86,024} = 0.233$$

The probability is 0.233 that a worker is a member of a labor organization and that he is in a non-agricultural industry.

Note that this probability can also be computed directly from the data in the Table. The conditional probability is

$$P(A_1|B_1) = \frac{20{,}044}{86{,}024} = 0.233 \text{ ["|}B_1\text{" means only } B_1 \text{ column is relevant].}$$

The answer is the same as computed by using the formula for conditional probability.

BAYES' THEOREM

In its general form, Bayes' theorem deals with specific events, such as A_1, A_2,...., A_k, that have prior probabilities. These events are mutually exclusive events that cover the entire sample space. Each prior probability is already known to the decision maker, and these probabilities have the following form, $P(A_1)$, $P(A_2)$,...., $P(A_k)$. The events with prior probabilities produce, cause, or precede another event, say B. A conditional probability relation exists between events A_1, A_2, ..., A_k and event B. The conditional probabilities are $P(B|A_1)$, $P(B|A_2)$, ..., $P(B|A_k)$.

Bayes' formula allows us to calculate the probability of an event, say, A_1 occurring given that event B has already occurred with a known probability, P(B). The probability of A_i occurring given that B has already occurred is the **posterior** (or revised) **probability**. It is denoted by $P(A_1|B)$. Thus, we are given $P(A_1)$ and the $P(B|A_1)$ which we use to calculate $P(A_1|B)$.

For any event A_i, Bayes' theorem has the formula

$$P(A_i \mid B) = \frac{P(A_i \text{ and } B)}{P(B)}$$

The probability that A_1 and B occur simultaneously is equal to the probability that A_1 occurs multiplied by the probability that B occurs given A_1. Thus, we have

$$P(A_1 \text{ and } B) = P(A_1)\, P(B|A_1)$$

Since A_1, A_2, ... , A_k form a partition of the entire sample space when event B occurs, only one of the events in the partition occurs. Thus, we have

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

We already know that for any event A_i,

$$P(A_i \text{ and } B) = P(A_i)\, P(B|A_i)$$

When we substitute the formula for $P(A_i$ and B) into the equation for P(B) we obtain

$$P(B) = P(A_1)\, P(B|A_1) + P(A_2)\, P(B|A_2) + ... + P(A_k)\, P(B|A_k)$$

If we then substitute P(B) and P(Ai and B) into the conditional probability, i.e. $P(A|B) = \frac{P(A_i \text{ and } B)}{P(B)}$, we obtain the generalized version of Bayes' formula, which is shown in the box.

Bayes' Theorem

$$P(A_i \mid B) = \frac{P(A_i)\, P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + ... + P(A_k)P(B|A_k)}$$

Bayes' theorem can be better understood with the following example:

Suppose that a personnel manager wishes to hire one person from among a number of job applicants for a clerical position. The job to be filled is fairly simple. On the basis of past experience, the personnel manager feels that there is a 0.80 probability of an applicant being able to fill the position. This probability is the prior probability.

A personnel manager usually interviews or tests each applicant, rather than selecting one candidate randomly. This procedure supplies additional direct information about the applicant. In light of this additional information, the personnel manager may revise the prior probability about an applicant's chances of success or failure at the job. The revised probability is the posterior probability.

The terms prior and posterior refer to the time when information is collected. Before information is obtained, we have prior probabilities. Bayes' theorem provides a means of calculating posterior probabilities from prior probabilities. The next example illustrates the use of Bayes' theorem.

Example 16

A personnel manager feels that an applicant has the following chances of success and failure in a given job opening:

P(success) = 0.8

P(failure) = 0.2

These probabilities are prior probabilities. They are the probabilities of success and failure prior to obtaining specific information about an applicant.

During the testing and interviewing period, the manager assigns each applicant a rating of 1 (above average), 2 (average), or 3 (below average). The records show that applicants who turn out to be successful have the following chances of receiving a 1, 2, or 3 rating:

P(1|success) = 0.7

P(2|success) = 0.2

P(3|success) = 0.1

The records also show the chances of unsuccessful candidates receiving a 1, 2, or 3 rating:

P(1|failure) = 0.1

P(2|failure) = 0.3

P(3|failure) = 0.6

These last six values are conditional probabilities. The value 0.6 represents the probability of receiving a 3 rating given that a candidate is a potential failure for the position.

The personnel manager will use Bayes' theorem to find successful candidates for each job on the basis of the applicant's rating.

In the above example, a candidate applying for the position can succeed with either a 1, 2 or 3 rating. Given a particular rating, the personnel manager must determine the probability of the applicant being successful in that position.

We use the multiplication rule to find the joint probability of an applicant having a rating of 1 and being successful in the position.

P(1 and success) = P(1|success) P(success)

= (0.7) (0.8) = 0.56

Computation of similar joint probabilities shows

P(1 and success) = 0.56 P(1 and failure) = 0.02

P(2 and success) = 0.16 P(2 and failure) = 0.06

P(3 and success) = 0.08 P(3 and failure) = 0.12

These results, in turn, allow us to compute the probability of an applicant receiving a 1, 2, or 3 rating during an interview. Use of the addition rule shows

P(1) = P(1 and success) + P(1 and failure)

= 0.56 + 0.02 = 0.58

P(2) = 0.22 and P(3) = 0.20

Suppose that a particular applicant is given a 1 rating. Bayes' theorem allows us to calculate the probability that the applicant will be successful. The posterior probability of an applicant being a success given that a 1 rating was assigned is calculated by the conditional probability formula as follows:

$$P(\text{success}|1) = \frac{P(1 \text{ and success})}{P(1)}$$

The formula says that the probability of a success given a 1 rating is equal to the probability of a 1 rating and a success divided by the probability of a 1 rating. By substitution, we have

$$P(\text{success}|1) = \frac{0.56}{0.58} = 0.97$$

Similarly, the posterior probability of a failure given a 1 rating is

$$P(\text{failure}|1) = \frac{P(1 \text{ and failure})}{P(1)} = \frac{0.02}{0.58} = 0.03$$

Thus, when an applicant has a 1 rating, the probability of being a potential success increases from 0.8 (the prior probability) to 0.97 (the posterior probability). So, the rating system appears to be reliable.

Similar computations show that the posterior probabilities for 2 ratings are:

$$P(\text{success}|2) = \frac{0.16}{0.22} = 0.73 \quad P(\text{failure}|2) = \frac{0.06}{0.22} = 0.27$$

For 3 ratings, the posterior probabilities are:

$$P(\text{success}|3) = \frac{0.08}{0.20} = 0.40 \quad P(\text{failure}|3) = \frac{0.12}{0.20} = 0.60$$

Remark

Bayes' theorem is a version of the multiplication rule that we studied previously, namely,

$P(A \text{ and } B) = P(A)\, P(B|A)$

or $P(A \text{ and } B) = P(B)\, P(A|B)$

Bayes' theorem is merely this rule in the form

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A)\, P(B|A)}{P(B)}$$

which is the formula we used to calculate posterior probabilities. Some preliminary computations, however, may be required to obtain P(A and B) and P(B).

Bayes' theorem allows us to revise prior probabilities in the light of new information. The results of the formula are posterior probabilities.

ADDITIONAL ILLUSTRATIONS

Illustration 1

A bag contains 5 white balls, 6 red balls and 7 blue balls.

a. A ball is drawn at random. Find the probability that it is
   i. White
   ii. Red

b. Two balls are drawn at random. Find the probability that
   i. Both are white
   ii. One is blue and the other is red.

c. Four balls are drawn at random. Find the probability that two are red and the other two are blue.

Solution

a. i. P(White) = $\frac{5}{18}$

ii. P(Red) = $\frac{6}{18} = \frac{1}{3}$

b. i. Probability that both the balls are white = $\frac{C(5,2)}{C(18,2)} = \frac{10}{153}$

ii. Probability that one ball is blue and the other one is red

$$= \frac{C(7,1)\times C(6,1)}{C(18,2)} = 7\times 6\times \frac{16!2!}{18!} = \frac{14}{51}$$

c. Probability that two are red and other two are blue is given by

$$\frac{C(6,2)\times C(7,2)}{C(8,4)} = \frac{\frac{6!}{2!.4!}\times\frac{7!}{2!.5!}}{\frac{18!}{4!.4!}}$$

$$= \frac{6!}{2!.4!}\times\frac{7!}{2!.5!}\times\frac{4!.4!}{18!}$$

$$= \frac{2\times 5\times 6\times 7}{15\times 16\times 17} = \frac{7}{68}$$

[**Note:** For all the above computations, we apply

$$[\ ] = [c(n,r)] = \frac{n!}{r!(n-r)!}$$

and n! = n(n – 1)(n – 2)……3,2,1].

Illustration 2

Medical records show that one out of ten individuals in a certain town has a low thyroid condition. If 20 persons in this town are randomly chosen and tested, what is the probability that at least one of them will have a low thyroid condition?

Solution

Let the event of a low thyroid condition be denoted as T

P(T) = $\frac{1}{10}$ therefore P($\overline{T}$) = 0.9

Probability that the selected 20 persons do not have a low thyroid condition

= $(0.9)^{20}$ = 0.1216

Probability that at least one of the twenty persons has a low thyroid condition

= 1 – 0.1216

= 0.8784

Illustration 3

The quality control department of a company has two machines A and B. Machine A is a new one and passes only 2% of the defective products. Machine B is an old one and due to some flaws passes 7% of the defective products. A product finds its way to finished goods stores only after it is passed by both the machines. What is the probability that

a. Machine A has passed a product, given that machine B has found it defective.

b. Machine B passes a defective product, given that machine A has passed it.

Solution

Let the event of the passing of a defective product by machine A be denoted by A. Let the event of the passing of a defective product by machine B be denoted by B.

A and B are independent events and the occurrence or non-occurrence of one event does not affect the probability of the occurrence of the other event.

a. The required probability is

$$P(A/\overline{B}) = P(A) = 0.02$$

b. The required probability is

$$P(B/A) = P(B) = 0.07$$

c. A defective product reaches the finished goods stores only if both the machines pass it. The required probability is

$$P(A \text{ and } B) = P(A) \,.\, P(B)$$
$$= (0.02) \,.\, (0.07)$$
$$= 0.0014.$$

Illustration 4

A bag contains 2 green balls and 4 red balls and a second bag contains 4 green balls and 3 red balls.

a. If a ball is drawn at random from one of the two bags, what is the probability that it is a green ball?

b. A ball is drawn at random from one of the bags and it turns out to be a green one. What is the probability that it came from the first bag?

Solution

Let the various events be denoted as follows:

B_1 – The event of the first bag being selected.

B_2 – The event of the second bag being selected.

G – The event of drawing a green ball.

R – The event of drawing a red ball.

$$(B_1) = P(B_2) = 0.50$$

$$(G/B_1) = \frac{2}{6}, P(G/B_2) = \frac{4}{7}$$

$$(R/B_1) = \frac{4}{6}, P(R/B_2) = \frac{3}{7}$$

a. Now,

$$(G) = P(G \text{ and } B_1) + P(G \text{ and } B_2)$$

$$(G \text{ and } B_2) = P(G/B_2)\ P(B_2)$$

$$= \frac{4}{7} \text{x } \frac{1}{2}$$

$$= \frac{4}{14}$$

$$P(G \text{ and } B_1) = P(G/B_1)\, P(B_1)$$

$$= \frac{2}{6} \text{ x } \frac{1}{2}$$

$$= \frac{1}{6}$$

$$P(G) = \frac{4}{14} + \frac{1}{6} = \frac{19}{42}$$

b. We need to find $P(B_1/G)$

We have from Bayes' theorem

$$P(B_1/G) = \frac{P(G/B_1)P(B_1)}{P(G/B_1)P(B_1)+P(G/B_2)P(B_2)}$$

$$P(B_1/G) = \frac{\frac{2}{6}\times\frac{1}{2}}{\frac{2}{6}\times\frac{1}{2}+\frac{4}{7}\times\frac{1}{2}} = \frac{\frac{1}{6}}{\frac{1}{6}+\frac{2}{7}}$$

$$= \frac{\frac{1}{6}}{\frac{19}{42}} = \frac{7}{19}$$

Illustration 5

Doctors at a hospital have discovered a new disease that runs in families. A company has discovered two drugs X and Y for the disease. The hospital has agreed to try the drugs on 400 persons. The 400 persons were selected in such a way that each has an 80 percent chance of getting the disease if given neither drug. In the group, 100 persons were administrated drug X, 100 persons were administrated drug Y and 200 persons were administrated both the drugs. Drug X reduces the probability of the disease occurrence by 35% and drug Y reduces the probability of the disease occurrence by 20%. The two drugs when taken together work independently. If a randomly selected person in the group gets the disease in the future, what is the probability that he was given both the drugs?

Solution

Let the various events be denoted as follows:

D – Events of getting the disease

X – Events of being administered the drug X

Y – Events of being administered the drug Y

Now,

$P(D/X) = (0.65)(0.80) = 0.52$

$P(D/Y) = (0.80)(0.80) = 0.64$

We need to get the probability P(X and Y)/D).

From Bayes' theorem we have

$$= \frac{P(D/(XandY)\,P(X\,and\,Y)}{P(D/X)P(X) + P(D/Y)P(Y) + P(D/(XandY)P(XandY)}$$

$$P(X \text{ and } Y) = \frac{200}{400} = 0.50$$

$$P(X) = \frac{100}{400} = 0.25$$

$$P(Y) = \frac{100}{400} = 0.25$$

$$P(X \text{ and } Y)/D) = \frac{(0.416)(0.50)}{(0.52)(0.25)+(0.64)(0.25)+(0.416)(0.50)}$$

$$= \frac{0.208}{0.498} = 0.4177$$

Illustration 6

In a firm, 40 percent of the work force are female, 25 percent of the female workers are management grade and 30 percent of the male workers are management grade. If one management grade worker is selected at random from this firm, what is the probability that the worker is a female?

Solution

Let the various events be denoted as follows:

A – Male worker

B – Female worker

M – Management grade workers

P(A) = 0.60 ; P(B) = 0.40 ; P(M/B) = 0.25 ; P(M/A) = 0.30

Required probability = P(B/M)

We have from Bayes' theorem

$$P(B/M) = \frac{P(M/B)\,P(B)}{P(M/B)P(B) + P(M/A)P(A)}$$

$$= \frac{(0.25)(0.40)}{(0.25)(0.40) + (0.30)(0.60)}$$

$$= \frac{10}{28} = \frac{5}{14}$$

Illustration 7

If 3 persons, selected at random, are stopped on a street, what are the probabilities that

a. All were born on Friday.

b. 2 were born on Friday and the other on Tuesday.

c. None was born on Monday.

Solution

The probability of being born on a particular day of the week is $\frac{1}{7}$

a. Required probability $= \frac{1}{7} \text{x } \frac{1}{7} \text{x } \frac{1}{7} = \frac{1}{343}$

b. If A, B and C are the persons selected, the events mentioned in the question could occur in 3 ways.

1. A and B were born on Friday and C was born on Tuesday – Event P
2. A and C were born on Friday and B was born on Tuesday – Event Q
3. B and C were born on Friday and A was born on Tuesday – Event R.

$$P(P) = \frac{1}{7} \text{x } \frac{1}{7} \text{x } \frac{1}{7} = \frac{1}{343}$$

$$P(Q) = \frac{1}{7} \text{x } \frac{1}{7} \text{x } \frac{1}{7} = \frac{1}{343}$$

$$P(R) = \frac{1}{7} \text{x } \frac{1}{7} \text{x } \frac{1}{7} = \frac{1}{343}$$

Required probability P(P or Q or R) = P(P) + P(Q) + P(R)

$$= \frac{1}{343} + \frac{1}{343} + \frac{1}{343} = \frac{3}{343}$$

Probability that none was born on Monday

$$= \frac{6}{7} \times \frac{6}{7} \times \frac{6}{7} = \frac{216}{343}$$

Illustration 8

A box of fuses contains 20 fuses, of which 5 are defective. If fuses are drawn one at a time from the box and not replaced, what is the probability that 3 draws will result in 3 defective fuses?

If the fuses are drawn one at a time and replaced immediately. What is the probability that the 3 draws will result in 3 defective fuses?

Solution

Let the various events be denoted as follows:

A – First draw results in a defective fuse.

B – Second draw results in a defective fuse.

C – Third draw results in a defective fuse.

a. $P(A) = \frac{5}{20}$

Since the fuses are not replaced, $P(B/A) = \frac{4}{19}$ and $P(C/A \cap B) = \frac{3}{18}$

$$P(A \cap B \cap C) = \frac{5}{20} \times \frac{4}{19} \times \frac{3}{18} = \frac{1}{114}$$

b. Since the fuses are replaced immediately after being drawn the events. A, B and C are independent and $P(A) = P(B) = P(C) = \frac{5}{20}$

$$P(A \cap B \cap C) = \frac{5}{20} \times \frac{5}{20} \times \frac{5}{20} = \frac{1}{64}.$$

Illustration 9

Mr. Krishna, a quality control manager of Gist Electric, questions the reliability of the two quality control checks in the food-processor manufacturing process. One check is performed by a worker who manually checks the processors and a second check is performed by a computer monitor. Krishna knows that 5% of the time, the worker is apt to miss a defective processor and that 2% of the time the computer will malfunction and fail to detect defective processors.

a. If Krishna finds that the computer was malfunctioning, what is the probability that the worker might have missed a defective processor?

b. If he knows that the worker missed a defective processor, what is the probability that he will find the computer had malfunctioned?

c. What is the probability that the worker will miss a defective processor and the computer will malfunction at the same time, allowing a defective processor to the factory?

Solution

Let A and B denote the events of a defective processor remaining undetected on account of (i) manual error, and (ii) computer malfunctioning respectively.

Given $P(A) = 0.05$, $P(B) = 0.02$ and that the events A, B are statistically independent, the required probabilities are given below:

a. Probability that the worker might have missed the defective processor given that the computer was malfunctioning.

$= P(A/B) = P(A) = 0.05$

b. Probability that the computer was malfunctioning given that the worker might have missed the defective processor.

$= P(B/A) = P(B) = 0.02$

c. Probability that the computer was malfunctioning and the worker might have missed the defective processor.

= P(AB) = P(A) x P(B)

= (0.05) x (0.02)

= 0.001

Illustration 10

There are two security analysts A and B. The analysis of A has been found to be accurate in 75% of the cases and the analysis of B is found to be accurate in 80% of the cases. If a particular security is given to them for analysis, what is the probability that they will contradict each other?

Solution

P(A) = 0.75 ; P(A') = 0.25

P(B) = 0.80 ; P(B') = 0.20

P(A' and B) = 0.25 x 0.80 = 0.20

P(B' and A) = 0.20 x 0.75 = 0.15

Probability that they will contradict each other

= 0.20 + 0.15

= 0.35 or 35%

Illustration 11

X and Y are two team leaders in a department in an organization. Of all the project proposals reaching the department head, 60% are proposed by X and 40% are proposed by Y. A proposal given by X has a 40% chance of being approved and a proposal given by Y has a 55% chance of being approved by the department head. You are required to find out

a. The probability of a project being approved.

b. The probability that a project has been proposed by X, given that the same has been approved.

Solution

Event	Probability	P(Project is approved/event)	P(Project is approved, Event)
Project X	0.60	0.40	0.60 x 0.40 = 0.24
Project Y	0.40	0.55	0.40 x 0.55 = 0.22
P (project is approved) = 0.46			

Probability of a project being approved = P (Project is approved, project is proposed by X) + P(Project is approved, project is proposed by Y)

= 0.24 + 0.22 = 0.46

P (Project is proposed by X/Project is approved)

$$= \frac{P(\text{Project is approved, project is proposed by X})}{P(\text{Project is approved})}$$

$$= \frac{0.24}{0.46} = 0.522.$$

SUMMARY

- Probability of an event is defined as the chance of its occurrence. According to the classical approach, probability is the ratio of the number of equally likely possible outcomes favorable for an event to the total number of possible outcomes.
- The Relative frequency definition of probability defines it as the proportion of times an event occurs, if the experiment is repeated several times under the same or similar conditions. In case of Subjective probability, the probability of an event is based on personal judgment.
- The chapter introduces to us the Venn diagram, which is a pictorial representation of the sample space of an experiment.
- Under the assumption that the elements of the sample space have an equally likely chance of occurring, the Marginal or Unconditional probability can be defined as the ratio of number of possible outcomes favorable to the event A to the number of possible outcomes.
- Conditional probability is the probability of occurrence of an event 'A', subject to the occurrence of a previous event, say B.
- If A and B are independent (i.e., the occurrence of A is no way related to the occurrence or non-occurrence of event B), then P(A/B) = P(A).
- If A and B are dependent, then P(A/B) = P(A and B)/P(B).
- Bayes Theorem deals with specific mutually exclusive events that have prior probabilities. It allows us to calculate the probability of an event, say A1, given that the event B has already occurred with a known probability, P(B). The probability P(A1/B) is called Posterior (or revised) probability.

Chapter XIV

Theoretical Distributions

After reading this chapter, you will be conversant with:

- Random Variable
- Expected Value
- Theoretical Distributions
- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- Additional Illustrations

Introduction

In the previous chapters, we studied the empirical or observed frequency distribution where, the actual data are collected, classified and tabulated in the form of frequency distribution. This data is usually based on sample studies and measured by statistical tools such as average, dispersion, skewness, correlation etc. Such a distribution gives not only the nature and form of sample data but also helps in formulating some ideas about the population characteristics. A more scientific way of drawing inferences about the characteristics of population is through theoretical distribution. The values of variables in the population distribution are distributed according to definite mathematically expressed probability law, which are based on 'Priori or a Posterior' inferences. Such corresponding probability distribution is known as Theoretical Probability Distribution. It enables us to fit a mathematical model for a given data.

In this chapter, we will study random variable, mathematical expectations, probability distribution functions, mean and variance in terms of probability functions, which provides tools for studying the theoretical distribution. We will also study the Univariate Probability Distributions – Binomial, Poisson and Normal Distribution. The Binomial and Poisson probability distribution are discrete probability distributions and the third Normal distribution is a continuous probability distribution.

RANDOM VARIABLE

A variable which assumes different numerical values as a result of random experiments or random occurrences is known as a random variable.

The rainfall measured in centimeters on each day of the monsoon season, the maximum temperature of each day for a city, the number of passengers traveling by train from Delhi to Mumbai everyday and the number of patients seen by a doctor each day are all examples of random variables. That is, the values assumed by these variables on each day would be random and cannot be accurately predicted.

If the random variable can assume any value within a given range, it is called a continuous random variable. On the other hand, if the random variable can assume only a limited number of values, it is called a discrete random variable. In examples cited in the previous para, rainfall and maximum temperature are examples of continuous random variables as they can register a wide variety of values within a given range. The number of persons traveling from Delhi to Mumbai everyday and the number of patients seen by a doctor each day are examples of discrete random variables as these values could only be whole numbers. You cannot have 353.5 persons traveling or 18.7 patients visiting the doctor.

EXPECTED VALUE

For taking decisions under conditions of uncertainty, the concept of expected value of a random variable is used. The expected value is the mean of a probability distribution. The mean is computed as the weighted average of the value that the random variable can assume. The probabilities assigned are used as weights. Thus, it is computed by summing up the random variables multiplied by their respective probabilities of occurrence.

$$E[X] = \Sigma X\, P(X)$$

Illustration 1

A person expects a gain of Rs.80, Rs.120, Rs.160 and Rs.20 by investing in a share. The probability distribution of the gains is as follows:

Gain (Rs.)	Probability
80	0.2
120	0.4
160	0.3
20	0.1

You are required to calculate the expected gain from the above data.

Solution

The expected gain from the share is,

(80 x 0.2) + (120 x 0.4) + (160 x 0.3) + (20 x 0.1)

= Rs.(16 + 48 + 48 + 2) = Rs.114

This expected value can be used to compare different investment opportunities. Suppose the investor could invest the amount in another security for which the probability distribution of gains is as follows:

Gain (Rs.)	Probability
150	0.1
80	0.8
20	0.1

The expected gain from the second security is,

(150 x 0.1) + (80 x 0.8) + (20 x 0.1)

= Rs.(15 + 64 + 2) = Rs.81

Since the expected gain from the second security is only Rs.81 as compared to Rs.114 from the first, the investor would do well to invest in the first security.

REMARKS

The points to be noted are:

i. The expected value calculation does not predict the value.

It does not mean that investment in the first security will always lead to a gain of Rs.114 and investment in the second security will always lead to a gain of Rs.81.

ii. Comparing the two expected values and taking a decision based on them only helps in ascertaining which of the alternatives is more likely to lead to higher profits.

Since the expected value of gain from the first security is higher than the expected value of gain from the second, one may conclude that the chance of higher gain is more likely from investing in the first rather than the second.

THEORETICAL DISTRIBUTIONS

Theoretical distribution is a distribution expected on the basis of previous experience or theoretical consideration. It is also called a probability distribution. Since the value of a random variable cannot be predicted accurately, by convention, probabilities are assigned to all the likely values that the variable may take. This is called Probability Distribution. For example, if the price of a share of Dowell Limited could have possible values of Rs.15, Rs.20, Rs.23, Rs.25

and Rs.30, the chances of the actual price taking on any of these values may be described by attaching probabilities of 0.12, 0.20, 0.08, 0.10 and 0.50 respectively to the prices. By enumerating the possible values and assigning probabilities specifically to each of these values, we are in fact, describing a probability distribution of share prices.

If a variable X can assume discrete set of values X_1, X_2, X_3 ….. X_k with probabilities P_1, P_2, P_3, ..… P_k respectively, where $P_1 + P_2 + P_3 + ….. P_k = 1$, such a distribution is known as Probability Distribution for X. The function P(x) which has the respective values P_1, P_2 ….. for $X = X_1, X_2, X_3, ….. X_k$ is called probability function or frequency function of X.

Probability function and relative frequency distribution are analog terms, with probability replacing relative frequencies. When the number of observations is large, the probability distribution is known as theoretical distribution or ideal limiting forms of relative frequency distribution. The probability distribution of random variable may be:

- Theoretical listing of outcomes and probabilities obtained from mathematical model.
- Empirical listing of outcome and its observed relative frequencies.
- Subjective listing of outcomes and their subjective or contrived probabilities.

In this chapter, we deal with first kind of probability distribution. To understand the difference between observed frequency distribution and theoretical distribution, let's take an example. If a coin is tossed, we expect to get 50 percent heads and 50 percent tails, if the experiment is carried out for longer period. Now if the coin is tossed for 100 times, we may get 40 tails and 60 heads. This is known as observation which differs from our expectation of 50 heads and 50 tails. This difference may be due to sampling fluctuations or due to fact of biased coin. At this point, it is necessary to know the expected behavior of the coin.

For discrete random variable, the probability distribution is exclusive listing of numerical outcomes for random variables so that the particular probability occurrence is associated with its outcome. A random variable is a numerical quantity whose value is determined by the random experiment. When this experiment is performed, its total outcome forms a set known as 'sample space' (S) of the experiment. The function of the variable random is known as 'probability function'. A random variable is of two types – discrete and continuous random variable. If the set of values defined by random variables over its sample space are finite, the random variable is said to be Discrete random variable. If the random variable assumes any real value in an interval, it is known as Continuous random variable. Let us now explain the relationship of these variables with probability distributions.

Discrete Uniform Distribution

If Y, the random variable is a discrete random variable, the probability function P(Y) is called Probability mass function and its distribution is known as 'Discrete Probability Distribution'. Let us now discuss this with an example. Acme Limited is a car manufacturer. The company can paint the car in 3 possible colors: White, Black and Blue. Until the population is sampled, the company does not know the demand for each color. Until such time, the company should assume that probability of demand for each color equals one-third.

If White, Black and Blue denote the events "Demand for White/Black/Blue car"

P(White) = P(Black) = P(Blue) = 1/3.

The Standard Discrete Uniform Distribution is obtained by representing events E_1, E_2, ….., E_K by the numbers 1, 2, ….., K.

(Remember, a random variable must assume numerical values.)

In the above example,

Event	X	P(X)
White	1	1/3
Black	2	1/3
Blue	3	1/3

It can be shown for the Standard Discrete Uniform Distribution

$\mu = E(X) = (K + 1)/2$

$\sigma^2 = V(X) = (K^2 - 1)/12$

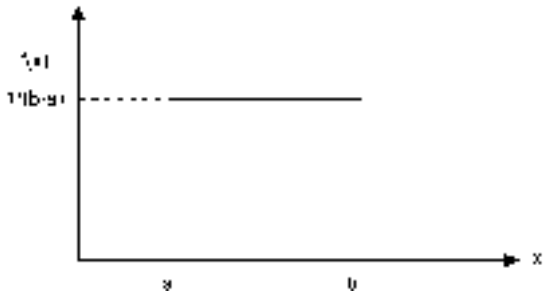
The examples of such distribution are Binomial distribution, Poisson distribution etc. These are discussed in the later pages of the chapter.

Continuous Uniform Distribution

If Y is a continuous variable its probability function is called 'probability density function' and distribution as 'continuous probability distribution'. Consider the interest earned on a bank deposit. Let X equal the value after the decimal point. (Assume no rounding off to the nearest paise.) For instance, if the interest = Rs.72.587, then X = 0.587.

X can take any value between 0 and up to but not including 1. We may assume that each value is equally likely. In which case, X is said to have a Continuous Uniform Distribution.

Figure 2



The probability density function of a typical random variable is given by

$$f(x) = \begin{cases} 1/(b-a) & a \le X \le b \\ 0 & \text{otherwise} \end{cases}$$

Note that the area under the curve equals 1.

It can be shown:

$\mu = E(X) = (a + b)/2$

$\sigma^2 = V(X) = (b - a)^2/12$

The example of such distribution is the normal distribution which is dealt later in the chapter.

BINOMIAL DISTRIBUTION

Binomial Distribution is associated with the name of a Swiss Mathematician James Bernoulli who discovered it in 1700 and was published first in 1713, eight years after his death. The binomial distribution is a probability distribution that expresses the probability of one set of dichotomous alternatives, i.e. success or failure. The process which gives rise to binomial distribution is known as Bernoulli trial or Bernoulli process. "A Bernoulli process is a process wherein an experiment is performed repeatedly, yielding either a 'success' or 'failure' in each

trial and where there is absolutely no pattern in the occurrence of successes and failures. The occurrence of a success or failure in a particular trial that does not affect and is not affected by the outcomes in any previous or subsequent trials. The trials are independent."

It is based on specific set of assumptions which involves the concept of a series of experimental trials. The assumptions of Bernoulli process are:

- The random experiment is performed repeatedly under same conditions for a fixed/finite number of times/trials, say n. In other words, n, the number of trials are fixed and finite.
- The outcome of random experiment has dichotomous classification of events. In other words, the outcome of each experiment has two mutually disjoint outcomes, i.e., Success (the occurrence of an event) and Failure (the non-occurrence of an event). Sample space S = { failure, success}.
- The result/outcome of any trial or sequence of trial is not affected by any preceding trial and does not affect the succeeding trial i.e., all trials are statistically independent.
- The probability of Success (S) (happening/occurrence of an event) of any trial denoted by p remains constant from trial to trial. The probability of failure (F) (non-happening or non-occurrence of an event) of any trial is denoted by 'q' = (1 – p) is also constant from trial-to-trial. We will not have binomial distribution if the probability of success is not the same.

For example, consider a batch of N light bulbs. Each bulb may be defective (S) or non-defective (F). The experiment involves selecting a light bulb and checking whether it is S or F. This experiment is called a Bernoulli Experiment since it has only two outcomes Success and Failure. Suppose it is known that there are M defective light bulbs in the batch. If we represent success by 1 and failure by 0, then

P (Success) = P (X = 1)= M/N = p (say)

P (Failure) = P (X = 0) = 1 – p = q (say)

X is said to be a random variable with Bernoulli distribution.

(Notice that a Bernoulli experiment can always be replicated by a (biased) coin with Head = 1, Tail = 0, P(1) = p)

Suppose the Bernoulli experiment is repeated n times under the same conditions. That is, after the light bulb is tested, it is put back into the batch. This way, the probabilities p and q remain unchanged. (This type of sampling is called Sampling with Replacement).

We sum up the Bernoulli Process as follows:

i. Each trial has only two possible outcomes i.e. p, the probability of success is constant for any trial and q = 1 – p is the probability of failure in any trial,

   In our example, the two possible outcomes are whether a bulb is defective or non-defective.

ii. The probability of the outcome of any trial remains fixed over time.

   In our example, the probability of the bulb being defective or non-defective remains fixed throughout.

iii. The trials are statistically independent.

In our example, the outcome of the bulb being defective or non-defective does not affect the outcome of any other bulb being so.

If X denotes the number of successes in n trials in order to satisfy the above assumptions, X is a random variable whose value ranges from 0,1,2,….,n ; since in n trials we may get no success/all failures, one success, two success, three

success…., and all n successes. We need to calculate the corresponding probabilities of 0,1,2,…n successes. The probability of r successes is given by:

$$p(r) = p(X=r) = {}^{n}C_r.p^r.q^{n-r};$$

Where,

r = 0,1,2,…..n.

Let X = Number of successes in n trials

Then,

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \; x = 0, 1, 2, ..., n \text{ where } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

In other words, if a coin is tossed for n times, the probability of various outcomes (0,1,2,…n), are given by successive terms of binomial expansion $(q+p)^n$ which is

$$(q+p)^n = q^n + {}^{n}C_1 q^{n-1}p + {}^{n}C_2 q^{n-2}p^2 + {}^{n}C_3 q^{n-3}p^3 + ... + {}^{n}C_r q^{n-r}p^r + ...p^n$$

These terms can be shown in the form of probability distribution table as follows:

Probability Table for Number of Tails

Number of Tails (X)	Probability P (X = r)
0	${}^{n}C_0 p^0 q^n = q^n$
1	${}^{n}C_1 q^{n-1}p$
2	${}^{n}C_2 q^{n-2}p^2$
3	${}^{n}C_3 q^{n-3}p^3$
r	${}^{n}C_r q^{n-r}p^r$
n	${}^{n}C_n p^n q^0 = p^n$

Thus, by expanding the binomial sum $(q + p)^n$, we can obtain the probabilities of 0,1,2,3,….n tails. Such a probability distribution is known as Binomial Probability Distribution or Binomial Distribution. The formula for Binomial probability distribution is

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

Where,

p = Probability of success in any trial,

q = 1 – p – Probability of failure in any trial,

n = Number of trials,

r = Number of Successes in n trials.

We can obtain the probable frequencies of various outcomes in N sets on n trials. Symbolically, it is represented as $N(q + p)^n$. It is calculated as follows:

$$N(q+p)^n = N[q^n + {}^{n}C_1 q^{n-1}p + {}^{n}C_2 q^{n-2}p^2 + {}^{n}C_3 q^{n-3}p^3 + ... + {}^{n}C_r q^{n-r}p^r + ...p^n$$

The frequencies obtained through the above formula is known as Expected Frequency or Theoretical Frequencies. Actual or Observed frequencies are those frequencies that are actually obtained through experience. So, there is difference between actual and observed frequency, but as N increases, the difference becomes small and small.

To find out the terms of expansion of $(q + p)^n$ and its coefficient, following rule should be remembered:

i. q^n is the first term.

ii. ${}^nC_1 q^{n-1} p$ is the second term.

iii. In each succeeding term, the power of p is increased by 1 and that of q is reduced by 1.

iv. The coefficient is calculated by multiplying the coefficient of preceding term by q's power in that term and dividing the obtained product by one more than power of p in preceding term.

Therefore, the expansion of $(q + p)^n$

$$(q+p)^n = q^n + {}^nC_1 q^{n-1} p + {}^nC_2 q^{n-2} p^2 + {}^nC_3 q^{n-3} p^3 + ... + {}^nC_r q^{n-r} p^r + ... p^n$$

Thus ${}^nC_1$, ${}^nC_2$, ${}^nC_3$, are called Binomial Coefficients. Thus, in expression $(q + p)^4$, we will have,

$$(q+p)^4 = q^4 + 4q^3 p + 6q^2 p^2 + 4q^1 p^3 + p^4$$

1,4,6,4 and 1 are coefficient. The above binomial distribution explains the following relationship:

- n +1 is the number of terms in a binomial distribution.
- The exponent of (p + q) is equal to n.
- The exponent of p and q are opposite to each other i.e the exponent of p are 0,1,2,3,…(n – 1), n and the exponent of q are n, (n – 1), (n – 2)……1, 0 respectively.
- The coefficient of the two central terms are identical when n + 1 is even. And when n is odd number, the coefficient of n + 1 is symmetrical ascending up to the middle of the series and then descending.

Pascal's triangle can also be used for calculating the coefficients of binomial distribution. Blaise Pascal was a famous French Philosopher and Mathematician who developed the theory of combinations as a triangular arrangement of numbers and to use this for calculating probability.

Figure 3: Pascal's Triangle Showing Coefficients of the Term$(Q + P)^n$

Row number No. of trails (n)	Binomial Coefficients																					No. of possible outcomes 2^n
1										1		1										2
2									1		2		1									4
3								1		3		3		1								8
4							1		4		6		4		1							16
5						1		5		10		10		5		1						32
6					1		6		15		20		15		6		1					64
7				1		7		12		35		35		21		7		1				128
8			1		8		28		56		70		56		28		8		1			256
9		1		9		36		84		126		126		84		36		9		1		512
10	1		10		45		120		210		252		210		120		45		10		1	1.024

Source: Gupta S.P 'Statistical Methods' Pgno:812.

Properties of the Binomial Distribution

i. The shape and location of binomial distribution changes with a change either in p or n i.e., either p changes for given change in n or n changes for a given change in p. The binomial distribution shifts to right when p changes for a given change in n.

ii. The binomial distribution's mode is equal to the value of x which has the largest probability. If np is an integer, mean and mode are equal. And if n is fixed, the mean and mode increase as p increases.

iii. The mean of the binomial distribution, np, increases as n increases with constant p.

iv. If n is large, neither p nor q is close to zero, the binomial distribution is approximated by a normal distribution with standardized variable given by $z = \frac{X - np}{\sqrt{npq}}$.

Importance of Binomial Distribution

The binomial distribution is a discrete probability distribution. It helps in describing variety of real life events. The binomial distribution can be used in following circumstances:

- The outcome of a trial is independent of the results of previous trial.
- The outcome of each trial will have one out of two possible outcomes.

Illustration 2

Find the probability of getting exactly three heads in 4 tosses of a biased coin, where

P(H) = 3/4 and P(T) = 1/4.

Solution

$P(X = 3) = (4c_3)\ (0.75)^3\ (0.25) = 4 \text{ x } (0.75)^3 \text{ x } (0.25)$

$= 0.421875$

Illustration 3

A coin is tossed five times. What is the probability of obtaining three or more heads?

Solution

In the case of unbiased coin, the probability of obtaining head and tail is equal i.e. p = q =1/2.

Here, n = 5 and the various possibilities of all the events are terms of expansion $(q + p)^5 = q^5 + 5q^4p + 10q^3p^2 + 10q^2p^3 + 5qp^4 + p^5$

The probability of obtaining 3 heads is $10q^2p^3 = 10 \times \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = \frac{10}{32} = 0.3125$

The probability of obtaining 4 heads is $5qp^4 = 5 \times \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^4 = \frac{5}{32} = 0.15625$

The probability of obtaining 5 heads is $p^5 = \left(\frac{1}{2}\right)^5 = 0.03125$

The probability of obtaining 3 or more heads = 0.3125 + 0.15625 + 0.03125 = 0.5

It can be shown for the Binomial Distribution

$\mu = E(x) = np$
$\sigma^2 = V(X) = npq$

Fitting of Binomial Distribution

Following procedure is followed for fitting a binomial distribution:

i. Calculate the values of p and q. If one of these values is known, the other can be calculated by the formula p = (1–q) and q = (1–p). The distribution is said to be symmetrical if the values of p and q are equal. If the values are not equal, the distribution is skewed. When p is less than 1/2, the distribution is positively skewed and the distribution is negatively skewed if p is more than $1/2$ . .

ii. Expand the binomial sum $(q + p)^n$. The power of n is always equal to one less than the number of terms in the expanded binomial. For example, if two coins are tossed, the binomial expansion will have three terms. Similarly, if 3 coins are tossed, there will be four terms and so on.

iii. Each term of the binomial expansion is multiplied by N (total frequency) so as to obtain the expected frequency in each category.

Illustration 4

8 coins are tossed at a time for 256 times. Find the expected frequency of successes if the number of heads is each throw are observed and recorded. Its result is given below. Also calculate the values of mean and standard deviation of the distribution.

No. of heads	Frequency	No. of heads	Frequency
0	2	5	56
1	6	6	32
2	30	7	10
3	52	8	1
4	67		

We are given n = 8 and N = 256. The probability of success (Heads) p = $1/2$ q = 1 – p = $1/2$.

Solution

We obtain the expected frequency of 1,2,3,….8 by expanding = $256\left(\frac{1}{2}+\frac{1}{2}\right)^8$

No. of Heads	Expected Frequency = $N \times {}^nC_r q^{n-r} p^r$
0	$256\left(\frac{1}{2}\right)^8 = 1$
1	$256 \times {}^8C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^7 = 8$
2	$256 \times {}^8C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^6 = 28$
3	$256 \times {}^8C_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^5 = 56$

No. of Heads	Expected Frequency = $N \times {}^{n}C_r q^{n-r} p^r$
4	$256 \times {}^{8}C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^4 = 70$
5	$256 \times {}^{8}C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^3 = 56$
6	$256 \times {}^{8}C_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^2 = 28$
7	$256 \times {}^{8}C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^1 = 8$
8	$256 \times \left(\frac{1}{2}\right)^8 = 1$
	Total = 256

The mean of the above distribution = $np = 8 \times \frac{1}{2} = 4$

The standard deviation = $\sqrt{npq} = \sqrt{8 \times \frac{1}{2} \times \frac{1}{2}} = 1.414$

Illustration 5

Fit a Binomial distribution to the following data:

X	:	0	1	2	3	4
F	:	28	62	46	10	4

Solution

In this problem, n = 4 ; N = 150

Mean of the distribution = $\bar{x} = \frac{\sum fx}{N} = \frac{0+62+92+30+16}{150} = \frac{200}{150} = \frac{4}{3}$

As per Binomial distribution mean = np, where n = 4 and the mean = 4/3

4 x p = 4/3 = p = 1/3 and q = 1- p = 2/3

Hence, the expected binomial distribution to be fitted the data is

$$150\left(\frac{2}{3}+\frac{1}{3}\right)^4$$

The theoretical frequencies for the binomial distribution are given below:

X	Theoretical Frequencies $N \times {}^{n}C_r q^{n-r} p^r$
0	150 x $(2/3)^4$ = 30
1	150 x $(2/3)^3$ x (1/3) = 59
2	150 x $(2/3)^2$ x$(1/3)^2$ = 44
3	150 x (2/3) x $(1/3)^3$ = 15
4	150 x $(1/3)^4$ = 2
	Total = 150

POISSON DISTRIBUTION

Poisson distribution is a discrete probability distribution developed by a French Mathematician Simeon D Poisson in 1837. It is widely used in statistical work. It is used in those cases where the chance of individual event being a success is less. Poisson distribution is used for describing the behavior of rare events such as number of accidents, printing mistakes in a book, etc., for this reason it is also called "the law of improbable events". It is obtained as a limiting case of Binomial Probability distribution when (i) the number of trials are large, (ii) the constant probability of success is small, (iii) np = mean is finite.

In the above three circumstances, the Binomial probability function tends to the probability function of the Poisson distribution. The Poisson distribution is defined as follows:

$$p(r)=\frac{e^{-m}m^{r}}{r!}$$

Where,

r = 0,1,2,3,……occurrence/success of an event

e = 2.7183 (the base of natural logarithms)

m = The mean of the Poisson distribution i.e. np.

- The variable X takes only integral values such as 0,1,2,3….∞ as the Poisson distribution is a discrete probability distribution.
- The probability at 0,1,2,3,…, successes can be obtained by putting r = 0,1,2,3,…., respectively i.e. the probabilities of 0,1,2,…. Successes, is given by successive terms of the expansion.

$$= e^{-m}\left[1+m+\frac{m^2}{2!}+\frac{m^3}{3!}....+\frac{m^r}{r}+.....\right]$$

This can be written in tabulated form:

No. of Successes (r)	Probability p(r)
0	$\frac{e^{-m}.m^0}{0!}=e^{-m}$
1	$\frac{e^{-m}.m}{1!}$
2	$\frac{e^{-m}.m^2}{2!}$
3	$\frac{e^{-m}.m^3}{3!}$
r	$\frac{e^{-m}.m^r}{r!}$

- The total probability is equal to one (q + p) = 1. p is very small in the case of Poisson distribution and q is almost equal to 1.

$$\sum_{r=0}^{\infty}p(r)=e^{-m}\left[1+m+\frac{m^2}{2!}+\frac{m^3}{3!}....+\frac{m^r}{r}+.....\right]$$

$$= e^{-m}\left[1+m+\frac{m^2}{2!}+\frac{m^3}{3!}....+\frac{m^r}{r}+.....\right]$$

$$= e^{-m}\times e^{m}$$

$$= e^{-m+m}$$

$$= e^{0}=1$$

The Poisson distribution has a single parameter 'm'. 'm' is called parameter because if we know 'm', we can determine all the probabilities of Poisson distribution. As 'm' increases, the distribution shifts to the right. Poisson probability distribution is also called Probability distribution of rare events because all probability distribution are skewed to the right.

Utility of Poisson Distribution

Poisson distribution is obtained as a limiting case of the Binomial distribution. It is used to explain the behavior of the discrete random variables where the probability of success is very small and the total number of cases are large. The Poisson distribution used in some practical situations are as follows:

1. It is used to study the number of defects per unit of manufactured product in Statistical quality control.
2. It is used to study the number of bacteria in biological sciences.
3. In insurance, it is used to study the number of casualities.
4. In waiting-time problem, it is used to study the number of telephone calls arriving or the number of incoming customers in super market.
5. In physical sciences, it is used to study the number of radio-active disintegrations of a radio-active element per unit.
6. Number of traffic arrivals at bus terminals, airports, docks, etc.
7. Number of suicides on a particular day or to determine the number of deaths or causalities due to rare disease in a given period in a particular district.
8. Number of accidents taking place on a busy road on a particular day.
9. Number of typographical errors per page in typed material.

Illustration 6

In certain district on an average 1 house in 500 has a fire during a year. If there are 1000 houses in that district, calculate the probability that exactly 5 houses will have fire during the year.

Solution

$\overline{X}$ = np

where,

n = 1000, p = 1/500

$$\overline{X} = 1000 \times \frac{1}{500} = 2$$

$$p(r) = \frac{e^{-m}.m^{r}}{r!},\ r = 0,1,2,3\ldots\ldots$$

$$p(5) = \frac{2.71823^{-2} \times 2^{5}}{5!} = \frac{.13534 \times 32}{120} = 0.036 \quad \text{(see table value of } e^{-m}\text{)}$$

Fitting of Poisson Distribution

Fitting of poisson distribution to a given frequency distribution is very simple. We have to compute the value of mean 'm' i.e., the average occurrence of the given distribution. Once 'm' is known, compute the frequency of 0 success and other probabilities of Poisson distribution can be calculated by using the following formula:

$$N(p_{x=x}) = N(p_{x-1}) \times \frac{m}{x}$$

$$N(p_0) = Ne^{-m}$$

$$N(p_1) = N(p_0) \times \frac{m}{1}$$

$$N(p_2) = N(p_1) \times \frac{m}{2}$$

$$N(p_3) = N(p_2) \times \frac{m}{3}$$

Illustration 7

Fit a poisson distribution to the followings data of mistakes per page in a book and also calculate the theoretical frequencies.

No. of mistakes per page	No. of times the mistake occurs
0	200
1	80
2	20
3	10
4	0

Solution

Calculation of Mean

X	f	fx
0	200	0
1	80	80
2	20	40
3	10	30
4	0	0
	N = 310	$\sum$ fx = 150

$$\overline{X} = \frac{\sum fx}{N} = \frac{150}{310} = 0.48$$

Mean of the distribution = m = 0.48 = 0.6188

Calculation of Expected Frequency

X	Expected Frequencies
0	$N(p_0) = .6188 \times 310 = 192$
1	$N(p_1) = N(p_0) \times \frac{m}{1} = 192 \text{ x } 0.48 = 92$
2	$N(p_2) = N(p_1) \times \frac{m}{2} = 92 \times \frac{0.48}{2} = 22$
3	$N(p_3) = N(p_2) \times \frac{m}{3} = 22 \times \frac{0.48}{3} = 3.5$
4	$N(p_3) = N(p_3) \times \frac{m}{4} = 3.5 \times \frac{0.48}{4} = 0.42$
	Total = 310

NORMAL DISTRIBUTION

Binomial and Poisson distributions are theoretical distributions useful for discrete variable, i.e., related to the occurrence of distinct and rare events. However, a continuous mathematical distribution is needed for those quantities whose magnitude is continuously variable. The most useful theoretical distribution for continuous variable is Normal distribution. The Normal distribution is also known as Normal Probability distribution. It is one of the most important continuous theoretical distribution in statistics. The statistical data relating to business and economic problem, social and physical sciences are displayed in the form of normal distribution. Therefore, normal distribution is considered a cornerstone of modern statistics.

Abraham De-moivre an English Mathematician was first to describe normal distribution as a limiting form of the binomial model in the year 1667-1754. In 1733, while dealing with the problem of game of chance, he formulated mathematical equation for this distribution. Gauss and Laplace rediscovered it in 1809 and 1812 respectively. Normal distribution is also known as Gaussian Distribution because this distribution was used for describing the theory of accidental errors of measurements involved in the calculation of orbit of heavenly bodies. The normal distribution is an approximation to binomial distribution. When the p and q are nearly equal, the normal approximation is good even when the value of n is small. When the p and q are not equal, even then the binomial distribution tends to the form of continuous curve (Normal distribution) only when the value of n is large as compared to the value of n required when p and q are equal. Thus, as the value of n increases, the normal approximation to the Binomial distribution is better and is a limiting case as $n \to \infty$. The limiting frequency curve obtained as n becomes large is known as the Normal Frequency Curve or Normal Curve. The normal curve can be represented in several forms. However, the basic form is related to mean μ and standard deviation σ . The formula is:

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)}{2\sigma^2}}$$

where,

X = Values of the continuous random variable

μ = Mean of the normal random variable

e = Mathematical constant approximated by 2.7183

π = Mathematical constant approximated by 3.1416

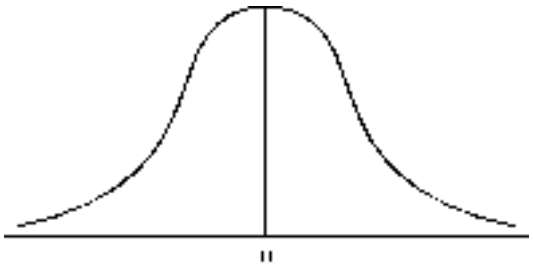
($\sqrt{2\pi}$ = 2.5066)

The ordinates obtained by the above formula is multiplied by N to get the ordinates of a particular distribution. So, the equation of normal curve relating to particular distribution is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$

The normal curve corresponding to the distribution with given total frequency N and the standard deviation σ will have maximum ordinate equal to $\frac{N}{\sigma\sqrt{2\pi}}$.

Figure 4



Source: Gupta S.P 'Statistical Methods' Pgno:838.

The normal distribution reflects the various values taken by many real life variables like the heights and weights of people or the marks of students in a large class. In all these cases, a large number of observations are found to be clustered around the mean value μ and their frequency drops sharply as we move away from the mean in either direction. For example, if the mean height of an adult in a city is 6 feet, a large number of adults will have heights around 6 feet. Relatively a few adults will have heights of 5 feet or 7 feet.

Further, if we draw samples of size n (where n is a fixed number over 30) from any population, the sample mean $\overline{X}$ will be (approximately) normally distributed with a mean equal to μ i.e. the mean of the population.

The characteristics of normal probability distribution with reference to the above figure are:

i. The curve has a single peak; thus it is unimodal. Since the distribution is unimodel, the only mode occurs at $X = \mu$.

ii. The mean of a normally distributed population lies at the center of its normal curve. It is the maximum point. If we move to either of the side from the mean, the curve declines but the curve never touches the base.

iii. Because of the symmetry of the normal probability distribution, the median and the mode of the distribution are also at the center.

iv. The two tails of the normal probability distribution extend indefinitely and never touch the horizontal axis.

v. The normal curve is 'bell-shaped' curve and is symmetrical in appearance. The two halves of the curve coincide. Therefore, the mean and median also coincide, i.e., the number of cases above the mean are equal to the number of cases below mean.

vi. The normal curve achieves its maximum height at its mean. Thus, in normal distribution mean and mode coincide. Hence, the mean, mode and median are equal in normal distribution.

vii. The inflexion point (Point at which curvature changes) occurs are $\overline{X} \pm \sigma$.

viii. The variables are continuous in normal distribution, whereas the variables are discrete in Binomial and Poisson distribution.

ix. Since the distribution is symmetrical, the quartiles (first and third) are equidistant from median.

x. Area property is the fundamental property of normal distribution. Under the normal curve, the area distributed are as follows:

   a. Mean $\pm 1\sigma$ covers 68.26% area of the observation i.e. 34.135% area of the observation lies on either side of the mean.

   b. Mean $\pm 2\sigma$ covers 95.45% or 0.9544 of the observation.

   c. Mean $\pm 3\sigma$ covers 0.9973 or 99.73% of the observation.

   The standard normal variate corresponding to X is

$$Z = \frac{X - \mu}{\sigma}$$

xi. The Mean Deviation (MD) about mean or median or mode is 4th or more precisely 0.7979 of the standard deviation.

Importance of Normal Distribution

Normal distribution occupies a central place in the theory of statistics. Its importance in statistical theory is as follows:

One of the most important application of the normal distribution has a remarkable property inherent in fundamental theorems of statistical theory i.e. Central Limit Theorem. The theorem is stated as,

"If $X_1, X_2, X_3 \ldots, X_n$ are n independent random variables following any distribution, under certain very general conditions, their sum $\sum X = X_1 + X_2 + \ldots + X_n$ is asymptotically normally distributed i.e. $\sum X$ follows normal distribution as $n \to \infty$."

An immediate consequence of this theorem is the following result.

"If $X_1, X_2, \ldots X_n$ is a random sample of size n from any population with mean μ and variance σ^2, the sample mean

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) = \frac{1}{n}\sum X$$

is asymptotically normal (as $n \to \infty$) with mean μ and variance σ^2/n".

In simple words, as the sample size n increases, the mean of the random sample approaches normal distribution.

- The normal distribution serves as a good approximation of discrete distribution such as Binomial and Poisson distribution as n, the number of trials becomes large. With large n, the computation of probability becomes tedious and time-consuming for discrete distribution. So in such cases, normal distribution can be used.
- In all probability distributions, we should expect that a standard normal variate lies between the limits $\pm$ 3 i.e., the value of standard normal variate will not go outside these limits.
- Normal distribution is used widely in Statistical Quality Control in industries for setting up of control limits.
- The sampling distributions such as Student's t-distribution, Fisher's Z-distribution Chi-square distribution conforms to normal distribution for large degrees of freedom. In fact, some of the distributions are based on the fundamental fact that the parent population from which the sample is drawn follows normal distribution.
- W. J. Youden, a contemporary Statistician of the National Bureau of Standard expresses the importance/admiration for Normal distribution artistically in the shape of normal curve as follows:

Figure 5

THE NORMAL

LAW OF ERROR

STANDS OUT IN THE

EXPERIENCE OF MANKIND

AS ONE OF THE BROADEST

GENERALISATION OF NATURAL

PHILOSOPHY. IT SERVES AS THE

GUIDING INSTRUMENT IN RESEARCHES

IN THE PHYSICAL AND SOCIAL SCIENCE AND

IN MEDICINE, AGRICULTURE AND ENGINEERING. IT IS AN

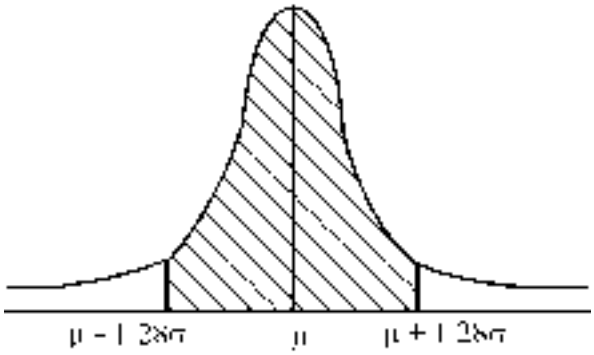
INDISPENSABLE TOOL FOR THE ANALYSIS AND THE INTERPRETATION

OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT.

Source: D. C. and Kapoor. V. K., Statistics (Theory, Methods and Application) Pgno:16.33.

If σ is the standard deviation of the normal distribution, 80% of the observation will be in the interval $\mu - 1.28\sigma$ to $\mu + 1.28\sigma$.

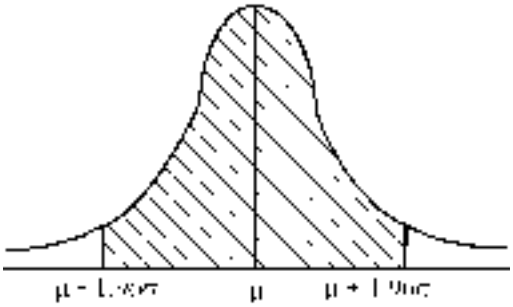
Figure 6



Source: Adapted from Gupta S.P 'Statistical Methods'.

395% of the observations will be in the interval $\mu - 1.96\sigma$ to $\mu + 1.96\sigma$.

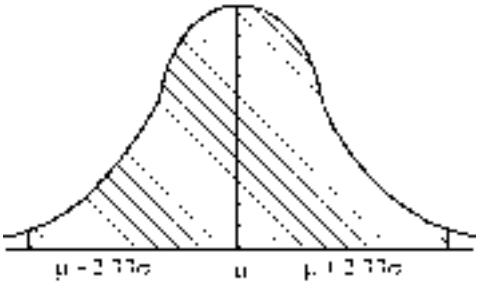
Figure 7



Source: Adapted from Gupta S.P 'Statistical Methods'.

98% of the observations will lie in the interval $\mu - 2.33\ \sigma$ to $\mu + 2.33\ \sigma$.

Figure 8



Source: Adapted from Gupta S.P 'Statistical Methods'.

Standard Normal Distribution

The Standard Normal Distribution is a normal distribution with a mean $\mu = 0$ and a standard deviation $\sigma = 1$. The observation values in a standard normal distribution are denoted by the letter Z.

Illustration 8

1. A population is normally distributed with mean = 0 and standard deviation = 1. What is the probability that an observation from the population will have a value between –1.28 and 1.28?

We know that for a normal distribution 80% of the observations lie between $\mu - 1.28\ \sigma$ and $\mu + 1.28\ \sigma$.

For a standard normal distribution $\mu = 0$ and $\sigma = 1$.

So, 80% of the observations will lie between –1.28 and + 1.28.

Hence, the probability that an observation will have a value between –1.28 and 1.28 is 80%.

2. What is the probability that an observation from a standard normal distribution will lie in the interval –1.96 to 1.96?

   95%.

3. What is the probability that an observation from a standard normal distribution will lie between –2.33 and + 2.33?

   98%.

STANDARDIZING NORMAL VARIABLES

Suppose we have a normal population. We can represent it by a normal variable X. Further, we can convert any value of X into a corresponding value Z of the standard normal variable, by using the formula $Z = \frac{X - \mu}{\sigma}$

Where,

X = the value of any random variable

μ = the mean of the distribution of the random variable

σ = the standard deviation of the distribution

Z = the ratio of difference between X and μ to the standard deviation is known as the Z score or standard score.

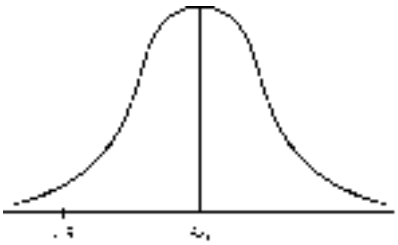
Illustration 9

1. A normal variable X has a mean of 56 and a standard deviation of 12.
2. A normal variable has a mean of 10 and a standard deviation of 5. What is the probability that the normal variable will take a value in the interval 0.2 to 19.8?

Solution

1. Find the Z value corresponding to the X value of –5.

$$Z = \frac{X - \mu}{\sigma} = \frac{-61}{12} = -5.08$$



2. Probability $(0.2 < X < 19.8)$

   = Probability $\left(\frac{0.2 - 10}{5} < Z < \frac{19.8 - 10}{5} \right)$

   = Probability $(-1.96 < Z < 1.96)$

   = 95%

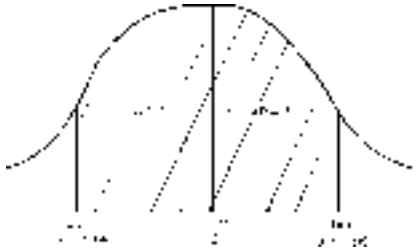
[because 95% of the area under the standard normal curve lies in the interval –1.96 to 1.96.]

We can see this from the Normal Tables as follows:

Area under the standard normal curve between 0 and 1.96 is 0.4750.

Due to symmetry of the standard normal distribution, area under the curve between –1.96 and + 1.96 is twice the area under the curve between 0 and + 1.96.

Probability (–1.96 < Z < + 1.96) = 0.95 or 95%



Note:

Any normal variable can be converted into a standard normal variable as illustrated above. Hence, we can use the standard normal distribution table to find the probability that the variable will take a value within any given interval.

ADDITIONAL ILLUSTRATIONS

Illustration 1

A random variable X has the following probability distribution.

X	0	1	2	3	4	5	6
P(x)	0.1	0.15	0.25	0.30	0.10	0.07	0.03

a. Find the expected value of X

b. Find the variance of X.

c. Find $P(1 \le X \le 5)$

d. Find $P(2 \le X)$

e. Find $P(X < 4)$.

Solution

The expected value or mean of the random variable X is given by

a. $E(X) = \sum XP(X)$

$= (0)(0.1) + (1)(0.15) + (2)(0.25)+(3)(0.3)+(4)(0.1) + (5)\ (0.07) + (6)\ (0.03) = 2.48$

b. The variance of X is given by $V(X) = \sum P(X - \overline{X})^2$

X	P(X)	$X - \overline{X}$	$(X - \overline{X})^2$	$P(X - \overline{X})^2$
0	0.10	–2.48	6.1504	0.6504
1	0.15	–1.48	2.1904	0.32856
2	0.25	–0.48	0.2304	0.0576
3	0.30	0.52	0.2704	0.08112
4	0.10	1.52	2.3104	0.23104
5	0.07	2.52	6.3504	0.444528
6	0.03	3.52	12.3904	0.371712
				2.1296

$V(X) = 2.1296$

c. $P(1 \le X \le 5)$ $= P(1) + P(2) + P(3) + P(4) + P(5)$
$= 0.15 + 0.25 + 0.30 + 0.10 + 0.07$
$= 0.87$

d. $P(X \ge 2)$ $= P(2) + P(3) + P(4) + P(5) + P(6)$
$= 0.25 + 0.30 + 0.10 + 0.07 + 0.03$
$= 0.75$

e. $P(X < 4)$ $= P(0) + P(1) + P(2) + P(3)$
$= 0.10 + 0.15 + 0.25 + 0.30$
$= 0.80$

Illustration 2

Fifteen management graduates cross all hurdles and reach the final phase of a selection process for the recruitment of management trainees by a company. The final phase is an interview and can gather from past data that, on an average, only 12 percent of the candidates qualifying for the interview get the job. Determine the following probabilities. (Assuming that there is no restriction on the number of management trainees to be recruited)

a. Probability that exactly 5 candidates get the job.

b. Probability that fewer than four candidates get the job.

Solution

In the event of a candidate getting the job could be called success, we have

$P = 0.12$, $q = 0.88$ and $n = 15$

The probability of r success in n trials is given by

$$P(r) = {}^{n}C_{r} q^{n-r} p^{r}$$

a. $P(5) = {}^{15}C_{5}(0.88)^{10}(0.12)^{5}$

$$= \frac{15!}{5!}(0.88)^{10}(0.12)^{5}$$

$= 0.0208$

b. $P(\text{Less than } 4) = P(0) + P(1) + P(2) + P(3)$

$= (0.88)^{15} + {}^{15}C_{1}(0.88)^{14}(0.12) + {}^{15}C_{2}(0.88)^{13}(0.12)^{2} + {}^{15}C_{3}(0.88)^{12}(0.12)^{3}$

$= 0.147 + 0.3006 + 0.287 + 0.1696$

$= 0.9042$

Illustration 3

On 25 questions, 4-answer multiple choice examination are taken by someone who knows absolutely nothing about the subject and must guess independently each answer,

a. What is the expected number of correct answers?

b. What is the probability of answering exactly 5 questions?

Solution

In the event of answering a question correctly is taken as a success, then we have

$p = 0.25$; $q = 0.75$ and $n = 25$

a. The mean is given by $np = (25)(0.25) = 6.25$

Therefore, the expected number of correct answers is 6.25

b. The probability of answering exactly r = 5 questions out of n = 25 questions correctly given by

$$P(r) = {}^{25}C_5(0.75)^{20}(0.25)^5 = 0.1645$$

Illustration 4

The owner of a bookshop has observed that the probability that a customer who is browsing through books will make a purchase is 0.3. Suppose that 15 customers browse through books in the section 'Probability and Statistics' each hour. Determine the following probabilities:

a. Probability that atleast one browsing customer will make a purchase during a specified hour.

b. Probability that no browsing customers will make any purchases during a specified hour.

c. Probability that no more than four browsing customers will make a purchase during a specified hour.

Solution

We can take the event of a customer making a purchase as a success.

Then we have p = 0.3, q = 0.7, and n = 15

The probability of r success in n trial is given by

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

a. P(at least 1) $= 1 - P(0)$

$= 1 - (0.7)^{15} = 0.9953$

b. $P(0) = 0.0047$

c. P(4 or less) $= P(0) + P(1) + P(2) + P(3) + P(4)$

$$= (0.7)^{15} + {}^{15}C_1(0.7)^{14}(0.3) + {}^{15}C_2(0.7)^{13}(0.3)^2 + {}^{15}C_3(0.7)^{12}(0.3)^3 + {}^{15}C_4(0.7)^{11}(0.3)^4$$

$= 0.5155$

Illustration 5

Out of 800 families with 4 children's each, how many would be expected to have (a) 2 boys and 2 girls, (b) Atleast 1 boy, (c) No girls, (d) Atmost 2 girls.

Assume equal probabilities for boys and girls.

Solution

Let the probability of the girl child be taken as p. Then p = 0.50; q = 0.50 and n = 4

The probability of r success is given by

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

a. Probability that a family of 4 will consist of 2 boys and 2 girls is given by

$$P(2) = {}^{4}C_2(0.50)^2(0.50)^2$$

$= 0.375$

Number of families expected to have two boys and two girls = 0.375 x 800 = 300

a. P(at least 1 boy) $= 1 - P(\text{no boys})$

$= 1 - P(4)$

$= 1 - {}^{4}C_4(0.50)^4(0.50)^0$

$= 1 - 0.0625$

$= 0.9375$

Number of families expected to have at least one boy = (0.9375)(800) = 750

b. P(No. girls) = P(0) = $(0.50)^4 = 0.0625$

Number of families expected to have no girls = (0.0625)(800) = 50

c. P(2 or less) = P(0) + P(1) + P(2)

$= {}^4C_4(0.50)^4(0.50)^0 + {}^4C_1(0.50)^3(0.50) + {}^4C_2(0.50)^2(0.50)^2$

$= 0.6875$

Number of families expected to have at the most 2 girls = (0.6875)(800) = 550.

Illustration 6

The marks of 100 students are normally distributed with mean 82 and standard deviation 3. How many students have got marks (a) equal to 82, (b)greater than 86, (c)between 79 and 85 inclusive and (d) less than 79? (Assume that the marks have been rounded off to the nearest integer).

Solution

Let X denote the random variable marks got by students.

a. The marks of 82 can have any value from 81.5 to 82.5. Therefore, the value of

$$Z = \frac{X-\mu}{\sigma} = \frac{81.5-82}{3} = \frac{-0.50}{3} = -0.17$$

$$Z = \frac{X-\mu}{\sigma} = \frac{82.5-82}{3} = \frac{0.50}{3} = 0.17$$

$P(81.5 < X\ 82.5) = P(-0.17 < Z < 0.17)$

$= 2(0.0675) = 0.135$

b. The marks of 86 can be considered to be 85.5 to 86.5 marks

$$Z = \frac{X-\mu}{\sigma} = \frac{86.5-82}{3} = \frac{4.50}{3} = 1.5$$

$P(X > 86.5) = P(Z > 1.5)$

$= 0.50 - 0.4332$

$= 0.0668$

c. We have to find the probability of students getting marks between 78.5 and 85.5

$$Z = \frac{X-\mu}{\sigma} = \frac{78.5-82}{3} = \frac{-3.50}{3} = -1.167$$

$$Z = \frac{X-\mu}{\sigma} = \frac{85.5-82}{3} = \frac{3.50}{3} = 1.167$$

$P(78.5 < X < 85.5) = P(-1.167 < Z < 1.1.67)$

$= 2(0.379) = 0.758$

d. We have to find the probability of students getting less than 78.5 marks

$$Z = \frac{X-\mu}{\sigma} = \frac{78.5-82}{3} = \frac{-3.50}{3} = -1.167$$

$P(X < 78.5) = P(Z < -1.167)$

$= 0.50 - 0.379$

$= 0.121$

Illustration 7

A bank receives about 1460 application per year for a current account with overdraft facility. We can see from the past records that the probability of an application being approved is, 0.8 on an average. Find the probability that in a particular year.

a. Not more that 1180 applications are approved.

b. More than 1180 applications are approved.

Solution

Here, probability of an application being approved = p = 0.8

Mean μ = np = (0.80) (1460) = 1168

Variance σ^2 = npq = (1460)(0.80)(0.20) = 233.6

Standard Deviation $\sigma = \sqrt{npq} = \sqrt{233.6} = 15.284$

The number of applications approved in a year can be said to follow a normal distribution with mean 1168 and standard deviation = 15.284

Let X be the random variable, denoting the number of applications approved in a year.

X = 1180

$$Z = \frac{X - \mu}{\sigma} = \frac{1180 - 1168}{15.284} = \frac{12}{15.284} = 0.7851$$

a. $P(X \leq 1180) = P(Z \leq 0.79)$
$= 0.50 + 0.2852$
$= 0.7852$

b. $P(X \geq 1180) = P(Z \geq 0.79)$
$= 0.50 - 0.2852$
$= 0.2148$

Illustration 8

An auto finance company has a number of schemes of car finance to suit its customers. In almost all cases, the returns of the company's investment in car finance vary from 20% to 24% with a mean return of 22%. Suppose a customer walks in, what is the probability that a deal will be struck with him which will give the company a return between 21% to 23% on his account? Assume that the probability of striking the deal with the customer is 20%.

Solution

Assumed that the returns on various accounts are normally distributed with a mean μ = 22% and σ. As all returns lies between 20-24%, those will be the outer limit (3σ limit) of the normal distribution.

$\mu + 3\sigma = 24$

$22 + 3\sigma = 24$

$\sigma = 0.666$

Let p_1 indicates the probability of the return being between 21% to 23%.

Let p_2 indicates the probability of striking a deal.

Given that the deal is struck, the probability of returns falling between 21 and 23% (p_1) is shown in the shaded region.

We know $Z = \frac{X - \mu}{\sigma}$

$= \frac{23 - 22}{0.666}$

The shaded region in normal curve = 0.4332 [from Normal Table]

Therefore, the probability = $0.4332 \times 2 = 0.8664$

The probability of striking a deal with the customer $p_2 = 0.20$

Therefore, the probability of striking a deal and returns falling between 21% and 23%

$p_1 \times p_2 = 0.8664 \times 2$

$= 0.1733$ or 17.33%

Illustration 9

The Placement division of ICFAI is interested in finding out the annual earnings of CFA's numbering around 1500. How large a sample it should take in order to determine the mean annual earnings within plus and minus Rs.2,000 and 95 percent confidence level?

You may assume that the earnings of CFAs follow normal distribution with a standard deviation of Rs.5,000.

Solution

We are required to find the sample size so that the mean annual earnings of the CFAs should fall within plus or minus Rs.2,000 at a confidence level of 95%.

The Z-value for 95% level of significance is 1.96

$$1.96\sigma_x = \text{Rs.2,000}$$

$$\sigma_x = \frac{\text{Rs.2,000}}{1.96}$$

$$= \text{Rs.1,020.408}$$

The standard error σ_x is also given by the following formula

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

$$\text{Rs.1,020.408} = \frac{5000}{\sqrt{n}}$$

$$\sqrt{n} = \frac{5000}{1020.408}$$

$$n = 24.01 \text{ or } 25$$

The sample may contain 25 CFAs at random

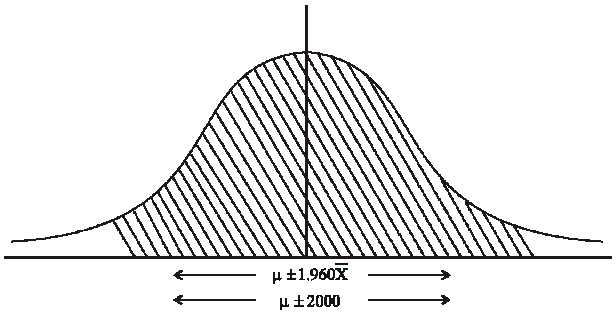


Illustration 10

Between the hours 1 P.M. and 3 P.M. the average number of phone calls per minute coming into the switchboard of a company is 2.35. Find the probability that during one particular minute, there will be at most 2 phone calls.

Solution

The number of calls per minute is denoted by X, and follows Poisson distribution with parameter m = 2.35. The probability of at most 2 phone calls during one particular minute is given as follows:

$$p(r) = \frac{e^{-m}.m^r}{r!},\ r = 0,1,2,3\ldots\ldots$$

$$P(r,2) = P(0) + P(1) + P(2)$$

$$= e^{-2.35}\left(1 + 2.35 + \frac{(2.35)^2}{2!}\right)$$

$$= 0.095374 \times (1 + 2.35 + 2.76125)$$

$$= 0.095374 \times 6.11125$$

$$= 0.5828543.$$

SUMMARY

- Probability distribution gives us an idea about the likely value of a random variable (a variable which assumes different numerical values as a result of random experiments) and the chances of occurrence of the various values. For taking decisions under uncertainty, the concept of expected value of a random variable is used where the expected value is the mean of a probability distribution.
- Given a probability distribution of paired data {x,y}, the covariance of the data measures the variability of a parameter (or data set) in relation to another parameter (or data set). Covariance can be used to find the variance of a sum of two or more random variables.
- In Discrete Uniform distribution, the random variables assume discrete values. Binomial distribution and Poisson distribution are examples of discrete uniform distribution. In the case of Continuous Uniform distribution, the random variable can take any value between zero and up to but not including 1. The Normal distribution is a continuous distribution in which a large number of observations cluster around the mean value 5 and their frequency drops sharply as we move away from the mean in either direction. The normal distribution is uni-modal and symmetrical, and standard normal distribution is a special case of normal distribution with mean = 0 and standard deviation = 1.
- In Binomial distribution the Bernoulli experiment (in which a variable has two outcomes: success (1) and failure (0)) is repeated n times with ‘p’ being the probability of success and ‘q’ being the probability of failure. The sampling in this case is done with replacement.
- Poisson distribution is obtained as a limiting case of the Binomial distribution. It is used to explain the behavior of the discrete random variables where the probability of success is very small and the total number of cases is large. It is used in various disciplines viz., business, life sciences etc.
- The importance of normal distribution in probability is found when the variable is continuous in nature. The normal distribution reflects the various values taken by many real life variables like the heights and weights of the people or the marks of students in a large class. In other words, it is of good use when the distribution is clustered.

Chapter XV

Linear Programming

After reading this chapter, you will be conversant with:

- Meaning of Linear Programming
- Review of Linear Functions
- The Graphical Method of Linear Programming
- The Simplex Method of Linear Programming
- Post Optimal Analysis
- Duality
- Additional Illustrations

MEANING OF LINEAR PROGRAMMING

Many business and economic situations are concerned with a problem of planning activity. In such cases, there are limited resources and the problem is to make such a use of these resources so as to yield the maximum production or to minimize the cost of production, or to give the maximum profit, etc. Such problems are referred as problems of constrained optimization. Linear Programming Problem (LPP) is a technique for determining the optimum schedule of interdependent activities in view of the available resources.

This LPP technique is designed to help managers in planning, decision-making and to allocate the resources. The management always tries to make the most effective use of organization resources. Resources include: machinery, labor, money, time, warehouse, space and raw materials. These resources may be used to produce services such as schedules for shipping, advertising policies and investment decisions etc.

All organizations have to make decisions about how to allocate their resources. There is no organization which operates permanently with unlimited resources. So, managements must continuously allocate scarce resources to achieve the organization's goals.

The word 'Programming' means 'Planning'. It refers to the process of determining a particular plan of action amongst several alternatives. The word 'Linear' indicates that all relationships involved in a particular problem are linear.

LPP is one of the most popular techniques of Operations Research. It is a mathematical technique for allotting limited resources of a firm in an optimum manner. The technique embraces almost every functional area of the business – production, finance, marketing, distribution etc. – in every type of industry. A wide variety of problems can be placed within the framework of this technique such as, to:

- decide on product quantities to maximize profit (product mix problems).
- determine the number of advertising units of different advertising media (Radio, T.V., Magazine) to ensure maximum exposures (media selection problems).
- find the quantity of components to be used in producing products at the minimum cost (alloy mix, fertilizer mix, food mix, paint mix, gasoline mix etc.)
- decide least-cost-route of transportation of units from different plants to different warehouses (transportation problems).
- establish optimal allocation of tasks to facilities (assignment problems).
- allocate order quantities, of an item, among different suppliers (assignment problems).
- select specific investments from among alternatives so as to maximize return, minimize risk (portfolio selection).
- develop a work schedule that follows a large restaurant, a hospital or a police station to meet staff needs at all hours with minimum number of employees (staffing problems).
- determine the most economic pattern and timings for flights so as to make the most efficient use of aircraft and crews (routing problems).
- find the combination of components to be produced from standardized raw material sized, (for example, paper, steel, and glass sheets in order to keep loss to minimum).
- select the shortest route for a salesman starting from a given city, visiting each of the specified cities and then returning back to the starting city (traveling salesman problems).

Terminology

The following terms are commonly used in the study of an LPP:

Decision Variable: Decision variables are the unknowns to be determined from the solution of an LPP model.

For example, decision variables in a product mix problem represent the quantities of the products to be produced, in a media selection problem they represent advertising units of different advertising media, in a diet mix problem they represent the quantities of different foods etc.

The essential requirements of the variable are:

- they should be inter-related in terms of consumption of resources.
- the relationships among the variables should be linear.

Constraint: A constraint represents a mathematical equation regarding limitations imposed by the problem characteristics. The constraints define the limits within which a solution to the problem must be found. Most often constraints represent the limits of a resource input. The constraints must be capable of being expressed in mathematical terms.

For example, assume that a company is manufacturing *x* and *y* numbers of two products '*A*' and '*B*' and each unit of these products requires respectively '*m*' hours and '*n*' hours of the machine shop capacity for which only '*t*' hours are available. Then the constraint on the production of product *A* and *B* may be expressed algebraically as:

$$m x + n y \leq t$$

Objective Function: An objective function represents the mathematical equation of the major goal of the system in terms of unknowns called decision variables. The objective function in linear programming is of optimization type – it can be maximizing profit function or minimizing cost function. The objective function, like constraints must be capable of being expressed in mathematical terms.

For example, if we assume that each unit of product *A* gives a profit of Rs.p_1 and each unit of product *B* gives profit of Rs.p_2, then the objective of the producer, say profit maximization, may be expressed in the mathematical form called objective function as under:

$$x\, p_1 + y\, p_2 = Maximum$$

The objective function is always non-negative. The coefficients associated with the variables in the objective function are constants and they represent either unit costs or unit profits of the items.

Linear Relationships: Linear programming deals with problems in which the objective function, as well as constraints, can be expressed as linear mathematical functions. Linear relationships have two properties: proportionality and divisibility.

Non-negativity Restrictions (Constraints): Essentially, the value of decision variables must be either zero or positive. Negative values of the decision variables imply negative production. Since, such a state in a real life situation is non-existent, decision variables must assume either zero or positive values. If *x* and *y* are decision variables, their non-negativity restrictions, shall be expressed as:

$$x \geq 0,\ \ y \geq 0$$

Feasible Solution: LPP helps to optimize a linear objective function subject to linear constraints of the variables. A set of values of the decision variables which satisfies all the constraints and non-negativity restriction is called feasible solution. There are usually a large number of feasible solutions to a problem.

Optimal Solution: A feasible solution which optimizes the objective function is called optimal solution. Optimal solution thus provides the best feasible choice of values which yields the highest (in case of maximization) or lowest (in case of minimization) value of the objective functions.

Stages of LPP

Each LPP involves three stages: (i) Problem Identification, (ii) Problem Formulation, and (iii) Problem Solving.

Problem Identification: Problem identification involves identification of the available alternatives, establishing the relationship between variables, specification of the constraints (i.e., available hours, space, materials, money etc.).

Problem Formulation: Problem formulation involves construction of a mathematical model from the given data. It requires identification of the decision variables, specifying the objective, setting up mathematical equations for the constraints and the objective and presenting the objective and constraints in a comprehensive form.

Problem Solving: Problem solving involves selection of the appropriate method, obtaining solution to the problem with the help of the selected method, and testing the solution for optimality.

An LPP may be solved either by the graphical method or by simplex method. Graphical method makes use of familiar graphical analysis and is used to solve a problem which involves two decision variables but any number of constraints. Simplex method is an iterative procedure where optimal solution to the problem is obtained from a series of arithmetical steps. The simplex method provides the means of solving complicated programming problems involving two or more decision variables.

Mathematical Formulation of LPP

Mathematical formulation of LPP consists of the following steps:

Step 1: Study the given problem and find the key decisions to be made (i.e., identify unknowns called decision variables).

Step 2: Identify the variables involved and denote them by symbols x_j $(j = 1, 2, ..., n)$.

Step 3: State the feasible alternatives (generally, $x_j \geq 0$, $\forall j$).

Step 4: Identify the constraints in the given problem and express them as linear inequations and/or equations.

Step 5: Identify the objective function and express it as a linear function of the decision variables.

Step 6: Express the objective function, constraints and non-negativity condition identified in steps 5, 4 and 3 in linear programming format.

General Formulation of LPP

The general formulation of the LPP can be stated as follows:

We find the values of *n* decision variables $x_1, x_2, ..., x_n$, to maximize or minimize the objective function

$$z = c_1x_1 + c_2x_2 + ... + c_nx_n \quad \ldots(1)$$

and also satisfy the *m*-constraints:

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + ... + a_{1j}x_j + ... + a_{1n}x_n \ (\le \text{ or } = \text{ or } \ge) b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2j}x_j + ... + a_{2n}x_n \ (\le \text{ or } = \text{ or } \ge) b_2 \\ \text{M} \qquad\qquad \text{M} \qquad\qquad \text{M} \\ a_{i1}x_1 + a_{i2}x_2 + ... + a_{ij}x_j + ... + a_{in}x_n \ (\le \text{ or } = \text{ or } \ge) b_i \\ \text{M} \qquad\qquad \text{M} \qquad\qquad \text{M} \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mj}x_j + ... + a_{mn}x_n \ (\le \text{ or } = \text{ or } \ge) b_m \end{array}\right\} \quad \ldots (2)$$

where constraints may be in the form of an inequality ($\le$ or $\ge$) or even in the form of an equation (=), and finally satisfy the non-negativity restrictions

$$x_1 \ge 0, x_2 \ge 0, ..., x_i \ge 0, ..., x_n \ge 0 . \quad \ldots (3)$$

The values of right side parameters $b_i \ (i = 1, 2, 3, ..., m)$ are restricted to non-negative values only.

- The linear function $z = c_1x_1 + c_2x_2 + ... + c_nx_n$ which is to be minimized (or maximized) is called the **objective function** of the general LPP.
- The inequations (2) are called the **constraints** of the general LPP.
- The set of inequations (3) is usually known as the set of **non-negative restrictions** of the general LPP.
- An *n*-tuple $(x_1, x_2, ..., x_n)$ of real numbers which satisfies the constraints of a general LPP is called a **solution** to the general LPP.
- Any solution to a general LPP, which also satisfies the non-negative restrictions of the problem, is called a **feasible solution**.
- Any feasible solution which optimizes (minimizes or maximizes) the objective function of a general LPP is called an **optimum solution**

REVIEW OF LINEAR FUNCTIONS

The function of a variable has been defined earlier in the chapter 'Functions and Calculus'. Linearity represents a special case of the relationship y = f(x). The relationship is defined as linear if, for all possible values of x and y, a given change in the value of x brings about a constant change in the value of y.

Example 1

Consider the following function,

y = 4 + 3x

The values of x and y and the changes in their values are given below:

x	Change in x	y	Change in y
–5	–	–11	–
–4	1	–8	3
–3	1	–5	3
–2	1	–2	3
–1	1	1	3
0	1	4	3
1	1	7	3
2	1	10	3
3	1	13	3

A plot of values of x and the corresponding values of y will trace a straight line.

The general expression of a linear function of one independent variable is:

y = a + bx

where,

x = independent variable

y = dependent variable

a = a numerical constant called 'intercept'

b = a numerical constant called 'slope'.

The expression for a linear function of n independent variables, x_1, x_2, x_3, …x_n, is

$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots\ldots + a_n x_n$

where,

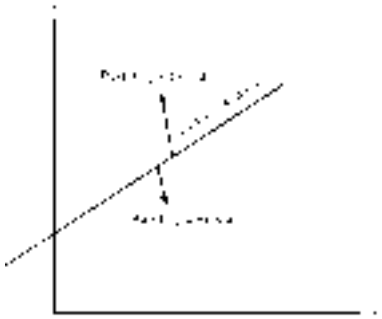
a_0, a_1, a_2, a_3, …… a_n are given numerical constants.

Some Important Results

The technique of solving linear programming problems is based on the study of the properties of linear functions.

Let us consider a linear function (Y) of one independent variable (X), y = a + bx, which is shown in the graph below.

Figure 1



In linear programming, we come across expressions of the following type:

$y - bx < a$

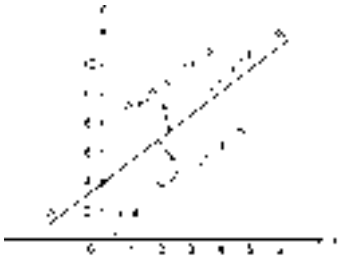
A line divides the plane into two parts. In one of the parts, the value of y – bx for all points in the part will certainly have a value less than the right-hand-side value 'a', and in the other part, the value of y – bx will be definitely greater than the value of 'a' for all the points in the part; whereas, for points on the line, the value of y – bx will be exactly equal to 'a'. We will be interested in the values of x and y for which the left hand side value is less than or equal to the right-hand side value 'a'.

Example 2

Consider the equation:

$y - 3x = 4$

This line is graphically represented below.



For all points on the line AB, the co-ordinates satisfy the equation:

$y - 3x = 4$

For any point below the line, i.e., Part I, we will have

$y - 3x < 4$

and for points above the line, i.e., Part II, we will have

$y - 3x > 4$

This can be easily verified by taking some points below and above the line.

In case of n number of variables x_1, x_2, x_3, ……,x_n, we get a graph in n-dimensional space with n co-ordinates. For any numerical values such as a_1, a_2, a_3.....,a_n and b, we define the equation for a plane as:

$a_1 x_1 + a_2 x_2 + \ldots\ldots + a_n x_n = b$

This plane divides the n-dimensional space into two parts. In one part, the co-ordinates of the points satisfy the inequality

$a_1 x_1 + a_2 x_2 + \ldots\ldots + a_n x_n < b$

and in the other part, the co-ordinates of the points satisfy the inequality

$a_1 x_1 + a_2 x_2 + \ldots\ldots + a_n x_n > b.$

GRAPHICAL METHOD

The graphical method is a simple one, and is the most easily understood of the several linear programming methods. A thorough knowledge of the graphical procedure provides necessary insight and confidence to understand the more advanced methods and concepts behind these methods. *However, it should be pointed out that the graphical method can be applied only in the case of two variables.* It cannot be applied to problems with many variables.

Maximization

Example 3

Consider a small foundry which specializes in the production of iron castings. For the sake of simplicity, assume that the foundry specializes in producing two types of castings – casting A and casting B. Because of a strong consumer demand for these products, it is assumed that the foundry can sell as many units as it produces. The profit is Rs.70 and Rs.40 for each of casting A and casting B respectively. The foundry manager should decide the quantity of these castings to be produced each week so as to maximize the total profit.

Production of castings requires certain resources like raw materials, labor and foundry capacity. The requirements and their availabilities are given in the following table:

Resources	Required per unit of		Available in a week
	Casting A	Casting B	
Raw material-1	2 kgs.	1 kg.	120 kgs.
Raw material-2	0.8 kgs.	none	40 kgs.
Labor	3 man-days	2 man-days	200 man-days
Foundry capacity	4 units	3 units	360 units

Let us formulate this problem in terms of mathematical equations or inequalities.

As the manager has to decide the number of type A and type B castings to be produced, let us define the variables:

q_1 = number of type A castings to be produced

q_2 = number of type B castings to be produced.

For this production schedule, the total profit will be

$70q_1 + 40q_2$

This function is known as the objective function which is to be maximized. If there are no constraints, the profit can be increased to infinity. In real life, there are restrictions of different kinds. These are formulated as constraints.

Let us consider raw material-1, of which only 120 kgs are available. If q_1 of type A castings and q_2 of type B castings are produced, then the requirement of raw material-1 is $2q_1 + 1q_2$, and this should be less than or equal to the available quantity of raw material-1. This can be shown by the following equation:

$$2q_1 + 1q_2 \leq 120$$

This implies that we are interested in the values of q_1 and q_2 for which the left-hand-side value is less than or equal to the right-hand-side value of 120. Otherwise, the requirement will exceed the availability and the production of that quantity will not be feasible.

By a similar argument, we get the constraints for raw material-2, labor and foundry capacity as:

Raw material-2	:	$0.8q_1 + 0q_2 \leq 40$
Labor	:	$3q_1 + 2q_2 \leq 200$
Foundry capacity	:	$4q_1 + 3q_2 < 360$

As one cannot produce negative quantities, we have the restrictions:

$$q_1 \geq 0,\ q_2 \geq 0$$

Putting together the above elements, the problem may be represented as:

Maximize Z: $70q_1 + 40q_2$ … (1)

Subject to constraints:

$$\begin{aligned} 2q_1 + q_2 &\leq 120 \\ 0.8q_1 + 0q_2 &\leq 40 \quad \text{… (2)} \\ 3q_1 + 2q_2 &\leq 200 \\ 4q_1 + 3q_2 &\leq 360 \end{aligned}$$

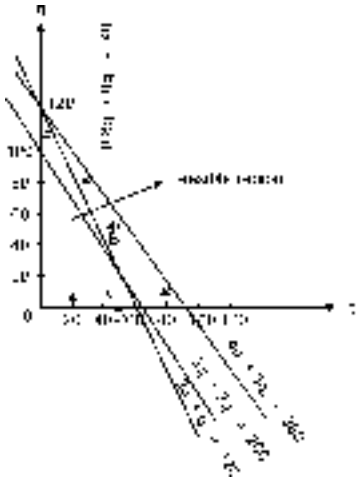
$$q_1 \geq 0,\ q_2 \geq 0 \quad \text{… (3)}$$

We have to find the values of q_1 and q_2 which will satisfy constraints (2) and (3) and at the same time maximize function (1). The function given in (1) is called an objective function. The inequalities in (2) are called constraints and the inequalities in (3) are called non-negativity restrictions or constraints. This problem cannot be solved by the calculus method because of the inequality constraints.

The first step in the graphical method of solution is to identify the region in the graph which corresponds to all pairs of values of q_1 and q_2 for which (2) and (3) are valid.

Let us consider the non-negativity restrictions given by (3). The values of q_1 and q_2 which satisfy these restrictions should fall in the first quadrant of the graph. Hence, We can ignore pairs of values of q_1 and q_2 which fall in other quadrants.

This is indicated by arrow marks on the q_1-axis (or x-axis) and q_2-axis (or y-axis) in the graph shown below.



Let us now find the region corresponding to the values of q_1 and q_2 for which the first constraint

$$2q_1 + q_2 \leq 120$$

is satisfied. To do this, we have to first draw the line

$$2q_1 + q_2 = 120$$

For this, we need to fix two points on this line. The points that we have chosen are:

$q_1 = 0$ $q_2 = 120$ and

$q_1 = 60$ $q_2 = 0$

By joining these points, we get the line, and the points below the line, indicated by arrows, will satisfy the first constraint.

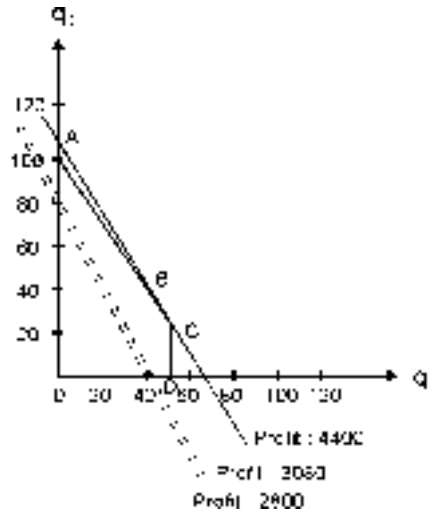
Other constraint equations are also drawn on the graph. The region common to all the regions identified gives the set of points for which the values of the co-ordinates q_1 and q_2 satisfy constraints (2) and (3). The region identified as OABCD is called the feasible region. It may be noted that all values of q_1 and q_2 which satisfy constraints (2) and (3), lie within the region OABCD and all points in the region OABCD will have the co-ordinates q_1 and q_2, which will satisfy constraints (2) and (3). Hence, an optimal solution to the problem should have co-ordinates q_1 and q_2 within or on the boundary of region OABCD.

Now, let us search for the optimal solution.

Suppose, we are interested in finding a product mix which will give a profit of say, an arbitrarily chosen value of Rs.2,800. To get the product mix, we have to search in the region OABCD to examine whether any point gives a profit of Rs.2,800. The easiest way is to draw the straight line whose equation is

$$70q_1 + 40q_2 = 2,800$$

and examine whether it passes through any points in the region OABCD. In the graph following, the feasible region OABCD and the above mentioned straight line are shown. We can observe that there are many points on this straight line which come under the feasible region, and each point will give the co-ordinates which refer to the production levels that yield the same profit of Rs.2,800. For example, take the two points (40,0) and (0,70) on this line. The production levels corresponding to these points are: (i) 40 of type A castings and 0 of type B castings, and (ii) 0 of type A castings and 70 of type B castings. It can be verified that each co-ordinate gives the same profit. Thus, the straight line drawn is also the profit line.



Suppose we wish to increase the profit, we look for a product mix which will give a profit of, say, Rs.3,080. As done earlier, we draw the line

$70q_1 + 40q_2 = 3,080$

and examine whether it passes through the region OABCD. This line is parallel to the first line and passes through the feasible region, thus indicating that it is possible to increase the profit to Rs.3,080. This suggests that, as we move up this line in the Northeastern direction, parallel to itself, we can obtain product mixes which will give higher and higher profits. We should move the line as far as possible without removing it completely from the region of feasible solutions as otherwise we will not find any feasible product mix which will satisfy the constraints. The optimal solution is then given by the point of final contact, which will be one of the corner points. In this case, the point is B, whose co-ordinates are (40, 40), indicating that the production level should be 40 for each of type B castings and this will yield a profit of

$70 \times 40 + 40 \times 40 = 2,800 + 1,600 = 4,400.$

REMARK

The following should be noted:

a. In some cases, the corner point may have co-ordinates which may be fractional. In cases where integer solutions are required (that is, some or all variables are required to take only integer values), other techniques called integer programming methods should be used. Otherwise, the solution which gives the fractional values, may be rounded off to arrive at a solution which is an approximation of the optimal solution.

b. If the objective function is parallel to an edge of the feasible region, then we get a number of product mixes, each of which will give the same maximum profit. This is a case of multiple optimal solutions. In the previous problem, consider the objective function to be $60q_1 + 40q_2$ instead of $70q_1 + 40q_2$. You may then ascertain that all the points on AB are optimal.

c. The optimal solution, if it exists, should occur at one of the corner points. These corner points are also called extreme points or vertices. Thus, to find the optimal solution to a linear programming problem, we need to search only the extreme points. It can be seen that there will be only a finite number of extreme points in any given problem.

Minimization

Example 4

A farmer is advised to utilize at least 900 kg of mineral A and 1200 kg of mineral B to increase the productivity of crops in his fields. Two fertilizers, F_1 and F_2 are available at a cost of Rs.60 and Rs.80 per bag. If one bag of F_1 contains 20 kg of mineral A and 40 kg of mineral B, and one bag of F_2 contains 30 kg each of mineral A and B, then how many bags of F_1 and F_2 should the farmers use to fulfill the requirement of both the types of minerals at an optimum low cost.

Let us formulate this problem in terms of mathematical equations or inequations. As the farmer has to decide on the number of bags of fertilizers F_1 and F_2, the variables may be defined as:

q_1 = number of bags of F_1

q_2 = number of bags of F_2

The objective function is minimization, that is, cost reduction. Here, the total cost is $60q_1 + 80q_2$. The restriction is that at least 900 kg of mineral A and 1200 kg of mineral B is required. Hence, we get the following constraints:

$20q_1 + 30q_2 \geq 900$ — requirement for mineral A

$40q_1 + 30q_2 \geq 1200$ — requirement for mineral B

As we cannot have negative quantities, $q_1 \geq 0$ and $q_2 \geq 0$,

the problem may be represented as

Minimize Z : $60q_1 + 80q_2$ … (1)

Subject to constraints:

$20q_1 + 30q_2 \geq 900$

$40q_1 + 30q_2 \geq 1200$ … (2)

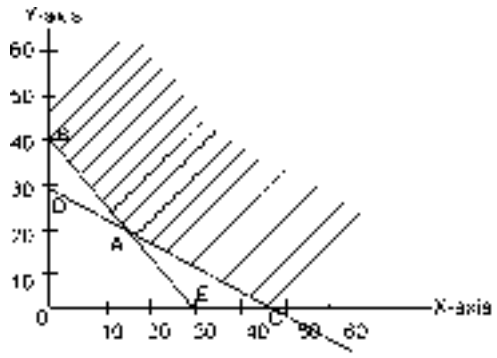
$q_1 \geq 0, q_2 \geq 0$ … (3)

We have to find the values of q_1 and q_2 which will satisfy constraints (2) and (3) and at the same time, minimize function (1).

After formulating the problem, each inequality is converted to an equality. Then, any arbitrary value (say, 0) is assigned to one variable in the equation and the corresponding value of the other variable is found. Consider the constraints which are written as equalities: $20q_1 + 30q_2 = 900$. If $q_1 = 0$, we get $q_2 = 30$ and if $q_2 = 0$, we have $q_1 = 45$. These two points are now plotted on a graph with q_1 on X-axis and q_2 on Y-axis. Joining the two points (0, 30) and (45, 0), we get a straight line corresponding to the above equation.

Consider the equation: $40q_1 + 30q_2 = 1200$. If we take $q_1 = 0$, then $q_2 = 40$, and if $q_2 = 0$, then $q_1 = 30$. Joining the two points (0,40) and (30,0), we get another straight line corresponding to the above equation. The next step is to graph the feasible region which satisfies all the constraints. For this, we should take the co-ordinates of the point of origin (0,0) and substitute in each inequality. If the statement is found to be true, shade the region towards the origin or else shade the region away from the origin.

Take the constraint $20q_1 + 30q_2 \geq 900$. If we substitute (0,0), we get (20 x 0) + (30 x 0) a $\not\geq$ 900. Since the statement is not true, we shade the region away from the origin. Similarly, for constraint $40q_1 + 30q_2 \geq 1200$, we shade the region away from the origin.



The region which satisfies all the constraints is the feasible region. Here, the region above ABC (that is, the intersection of all shaded regions) is the feasible region. Now, we should compute the co-ordinates of the corner points B, A and C of the feasible region. We know that the co-ordinates of B are (0,40) and that of C are (45,0). For point A, which is an intersection of the two straight lines of equations $20q_1 + 30q_2 = 900$ and $40q_1 + 30q_2 = 1200$, we find the co-ordinates by solving the simultaneous equations

$$20q_1 + 30q_2 = 900 \quad \ldots (1)$$

$$40q_1 + 30q_2 = 1200 \quad \ldots (2)$$

Subtracting equation (1) from (2) we get $20q_1 = 300$. Therefore, $q_1 = 15$ and $q_2 = 20$. Hence, the co-ordinates of A are (15,20).

The next step is to substitute the co-ordinates of the corner points of the feasible region in the objective function and choose the optimal solution (that is, the values that give the lowest cost).

We thus get the following volumes:

At A (15, 20), $Z = 15 \times 60 + 20 \times 80 = 2500$,

B (0, 40), $Z = 0 \times 60 + 40 \times 80 = 3200$, and

C (45, 0), $Z = 45 \times 60 + 0 \times 80 = 2700$.

From the above calculations, we find that Z assumes a minimum value at A (15,20). Therefore, the optimal value of $q_1 = 15$ and $q_2 = 20$. Hence, the farmer should buy 15 bags of fertilizer F_1 and 20 of fertilizer F_2 in order to meet the optimal requirements.

THE SIMPLEX METHOD

In large sized linear programming problems, the solution cannot be obtained by the graphical method and hence a more systematic method has to be developed to find the optimal solution. The "Simplex Method" developed by George B. Dantzig is an efficient algorithm to solve such problems. The simplex method is an iterative procedure for moving from one extreme point with a low profit value to another with a higher profit value until the maximum value of the objective function is achieved.

An application of the simplex method is illustrated below with the problem solved by the graphical method.

Maximization

Example 5

Maximize Z: $70q_1 + 40q_2$

Subject to:

$$2q_1 + q_2 \leq 120$$
$$0.8q_1 + 0q_2 \leq 40$$
$$3q_1 + 2q_2 \leq 200$$
$$4q_1 + 3q_2 \leq 360$$
$$q_1 \geq 0, q_2 \geq 0$$

The first step in applying the simplex method is to convert all inequalities into equations. This conversion could be accomplished by utilizing the concept of slack or unused resources. If we define:

S_1 = slack raw material-1

S_2 = slack raw material-2

S_L = slack labor

S_C = slack foundry capacity

Then, the constraints will be:

$$2q_1 + q_2 + S_1 = 120$$
$$0.8q_1 + 0q_2 + S_2 = 40$$
$$3q_1 + 2q_2 + S_L = 200$$
$$4q_1 + 3q_2 + S_C = 360$$
$$q_1 \geq 0,\ q_2 \geq 0,\ S_1 \geq 0$$
$$S_2 \geq 0,\ S_L \geq 0,\ S_C \geq 0$$

The profit contribution of slacks is taken as 0 so that the objective function is

$$70q_1 + 40q_2 + 0S_1 + 0S_2 + 0S_L + 0S_C$$

which is to be maximized. Let us call the original variables q_1 and q_2 as regular variables and the others as slack variables.

We must first form an initial solution to the constraints. This is obtained by assigning the value '0' to the regular variables q_1 and q_2 i.e., we shall start at the point 0 in the graph. The values of slack variables will now be:

S_1 = 120

S_2 = 40

S_L = 200

S_C = 360

The non-flow variables are called basic variables and the others are called non-basic variables (Basic: S_1, S_2, S_L, S_C; Non-basic: q_1, q_2).

We shall now construct the initial table and update it eventually. The explanation is given at the end of each table.

Table 1

Profit			C_j	70	40	0	0	0	0
Variables				q_1	q_2	S_1	S_2	S_L	S_C
	Profit	Variable	Solution						
	1	2	3	4	5	6	7	8	9
	0	S_1	120	2.0	1.0	1	0	0	0
	0	S_2	40	(0.8)	0	0	1	0	0
	0	S_L	200	3.0	2.0	0	0	1	0
	0	S_C	360	4.0	3.0	0	0	0	1
	Z_j			0	0	0	0	0	0
		$(Z_j - C_j)$		–70	–40	0	0	0	0

* Profit = 0

EXPLANATION

1. Columns (4) to (9) represent all the variables that appear in the problem, that is, q_1, q_2, S_1, S_2, S_L and S_C.
2. First, fill in the second row with the variables.
3. Fill in the first row with the profit contribution of these variables. The profit contributions are the coefficients of the variables in the objective function.
4. Fill column (2) with the slack variables, S_1, S_2, S_L, and S_C.
5. Fill column (1) with the profit contribution of variables that appear in column (2).
6. Fill column (3) with the solution values of the variables in column (2).
7. Column (4) is filled with coefficients of q_1 in the constraints. Similarly, columns (5) to (9) are filled with the coefficients of q_2, S_1, S_2, S_L and S_C respectively, in the constraints.
8. The value in the profit cell {column (3), last row} is obtained by multiplying the elements of column (1) with the elements of column (3), that is,

   $0 \times 120 + 0 \times 40 + 0 \times 200 + 0 \times 360 = 0$

   This implies that, for the initial solution $q_1 = 0$, and $q_2 = 0$ (that is, no production), the profit realized is 0.
9. The cells in the last row under columns (4) to (9) are called $(Z_j - C_j)$ cells.

   These values indicate the increase in the objective function per unit increase in the value of the variable currently at 0. In the initial table, this row is obtained by subtracting C_j from Z_j where Z_j is the summation of products of each value in columns (4) to (9) and the value in column (1). For instance, the Z_j value for column (4) is $2 \times 0 + 0.8 \times 0 + 3 \times 0 + 4 \times 0 = 0$. For column (5), Z_j is $1 \times 0 + 0 \times 0 + 2 \times 0 + 3 \times 0 = 0$.

Table 1 is now complete. At the end of each table, we should decide whether to update it to obtain a better solution, or to stop. This is done by the following steps:

Step 1: Is the solution indicated in the table optimal?

The answer to this question is obtained by looking at the $(Z_j - C_j)$ values in the last row. As mentioned above, these values indicate the profit that could be gained by increasing the production levels.

For example, the $Z_j - C_j$ value corresponding to variable q_1 is –70. This indicates that, by not producing type A castings, the foundry has lost Rs.70 per unit and hence by increasing the production level of type A castings, the foundry can increase its profit at the rate of Rs.70 per unit increase in production. Similarly, the foundry can increase its profit by Rs.40 by increasing the production level of type B castings. Thus, as there is scope for improving the profit, we have not reached the optimal solution.

Thus, we can answer the initial question by looking at all the $(Z_j - C_j)$ values. If all these values are greater than or equal to zero, it implies that we have reached the optimal solution and hence we should stop. If one of the $Z_j - C_j$ values is negative, then we should go to the next step.

Step 2: Find the variable to 'enter solution'.

This means identifying the product for which the production level has to be increased. We have seen that it is optimal to increase the production level of q_1. The rule is:

Identify the least negative $Z_j - C_j$ value. Thus, the corresponding (non-basic) variables will increase in value. In this case, q_1 has been selected to be a basic variable.

Table 2

Profit		C_j	70	40	0	0	0	0
Variables			q_1	q_2	S_1	S_2	S_L	S_C
Profit	Variables	Solution						
1	2	3	4	5	6	7	8	9
0	S_1	$120 - \frac{(2 \times 40)}{0.8} = 20$	0	$1.0 - \frac{(2 \times 0)}{0.8} = (1.0)$	$1.0 - \frac{(2 \times 0)}{0.8} = 1.0$	$0 - \frac{(2 \times 1.0)}{0.8} = -2.5$	$0 - \frac{(2 \times 0.0)}{0.8} = 0$	$0 - \frac{(2 \times 0)}{0.8} = 0$
70	q_1	$\frac{40}{0.8} = 50$	$\frac{0.8}{0.8} = 1$	$\frac{0}{0.8} = 0$	$\frac{0}{0.8} = 0$	$\frac{1}{0.8} = 1.25$	$\frac{0}{0.8} = 0$	$\frac{0}{0.8} = 0$
0	S_L	$200 - \frac{(3.0 \times 40)}{0.8} = 50$	0	$2.0 - \frac{(3 \times 0)}{0.8} = 2.0$	$0 - \frac{(3 \times 0)}{0.8} = 0$	$0 - \frac{(3 \times 1)}{0.8} = -3.75$	$1.0 - \frac{(3 \times 0)}{0.8} = 1.0$	$0 - \frac{(3 \times 0)}{0.8} = 0$
0	S_C	$360 - \frac{(4 \times 40)}{0.8} = 160$	0	$3.0 - \frac{(4 \times 0)}{0.8} = 3.0$	$0 - \frac{(4 \times 0)}{0.8} = 0$	$0 - \frac{(4 \times 1)}{0.8} = -5$	$0 - \frac{(4 \times 0)}{0.8} = 0$	$1 - \frac{(4 \times 0)}{0.8} = 1.0$
	Z_j		70	0	0	87.5	0	0
	$Z_j - C_j$		0	–40	0	87.5	0	0
* Profit = 70 x 50 = 3500								

Step 3: Find the variable to 'leave solution'

This is obtained by answering the following question:

In Step 2, let us decide to increase the value of q_1, that is, increase the production level of type A castings. By how much can the production of type A castings be increased?

The answer is; Increase the production until one of the resources gets exhausted. The requirement per unit of q_1 is given in Column (4) of the table. If the value is positive, it means that there is a positive requirement. If the value is 0 or negative, it means that particular resources are not required for increasing the value of q_1. By applying this simple logic, it is observed that the minimum positive ratio of the elements of Column (3) with the elements of the column selected in Step 2, that is, column (4) will indicate the resource which will be the first to be exhausted.

$$\text{Min}\left[\frac{120}{2.0}, \frac{40}{0.8}, \frac{200}{3.0}, \frac{360}{4.0}\right]$$

The minimum value is 50 corresponding to the ratio 40/0.8 and the corresponding variable is S_2, that is, the resource raw material 2, gets exhausted by producing 50 units of type A castings. We cannot increase the value of q_1 beyond 50 at this stage. As the value of S_2 decreases to 0, we replace S_2 by q_1 in Column (2). We then identify the row corresponding to S_2.

Step 4: Identify the pivot elements

This is the one element common to the column identified in Step 2 and the row identified in Step 3, that is, 0.8. This element is circled in the table. The column and row where the pivot element lies are called pivot column and pivot row respectively.

We are now ready to update Table 1 and construct Table 2. The format of Table 2 is the same as Table 1. The rules for updating are explained below in Table 2.

EXPLANATION

1. Fill in Row 1 and Row 2 of Tableau 2 in the same way as Tableau 1. Fill in Column (2) by replacing the variable selected in Step 3 with the variable selected in Step 2.
2. Fill in Column (1) with the profit contribution of the variables in Column 2.
3. Update the pivot row by dividing each element by the pivot element.
4. Update the pivot column by filling it with zeros. Note that the value of the key element in an updated table will be 1 and the values of the key or pivot column will be zero. Also, if a certain value in the pivot column is 0, then all the values in the corresponding row of the updated table remain the same as in the previous table. If the key row contains a 0, then the values of the corresponding columns in the updated table will remain the same as in the previous table. For example, we have 0 in the key row of Table 1 which corresponds to columns 5, 6, 8 and 9. One can notice that in Table 2, these columns retain the values of Table 1.
5. Update all other elements as follows:

   The first element in column 3 is 120 and it is to be updated. From this element we move to the pivot column, then to the pivot and then to the column we started with and trace the elements. These are:

   Value to be updated 120 2.0

   (40) 0.8 Pivot

   $$\text{The Updated value} = \text{Old value} - \frac{(\text{Product of the diagonal element})}{\text{Pivot element}}$$

or updated value =

$$\text{Old value} - \left(\frac{\text{Corresponding value in the previous key row} \times \text{Corresponding old row no. in previous key column}}{\text{Pivot element}} \right)$$

$$= 120 - \frac{(2.0 \times 40)}{0.8} = 20$$

Similarly, the third, fourth and the profit values in column 3 are updated as follows:

Value to be updated: 40 (0.8) → Pivot new value $= 200 - \frac{(3.0 \times 40)}{0.8}$

200 3.0 $= 50$

Value to be updated: 40 (0.8) → Pivot new value $= 360 - \frac{(4.0 \times 40)}{0.8}$

360 4.0 $= 160$

Value to be updated: 40 (0.8) → Pivot new value $= 0 - \frac{(-70) \times 40}{0.8}$

0 –70 $= 3{,}500$

The other elements are also updated in a similar fashion. The calculations are shown in the table.

The solution indicated in Table 2 is (from columns 2 and 3):

$S_1 = 20$

$q_1 = 50$

$S_L = 50$

$S_C = 120$

The variables q_2 and S_2 which do not appear in Column 2 take a value 0. Notice that the solution $q_1 = 50$, $q_2 = 0$ corresponds to the point D in the graph. Thus, we moved from point 0 to point D in the graph by updating Table 1.

At the end of Table 2, the four steps mentioned at the end of Table 1 are repeated. These are:

Step 1: *Is the solution indicated in Table 2 optimal?*

The $(Z_j - C_j)$ values are now (0, –40, 0, 87.5, 0, 0) and the negative values present imply that we have not yet reached the optimal solution.

Step 2: *Find the variable to enter the solution.*

The negative $(Z_j - C_j)$ value is –40 and this corresponds to the variable q_2. Hence, q_2 is the variable which enters the solution in the next iteration.

Step 3: *Find the variable to leave the solution.*

The ratios to be considered are: $\left[\frac{20}{1.0}, \frac{50}{2.0}, \frac{160}{3.0}\right]$.

The ratio corresponding to q_1 is ignored as the corresponding value in column 5 is 0. The minimum ratio is 20 corresponding to 20/1.0. Hence, the variable S_1 leaves the solution, and becomes non-basic.

Step 4: *Identify the pivot element.*

From steps 3 and 4, we find that the pivot element is 1.0, the first element in column 5, and this is circled.

We are now ready to construct Table 3. Table 2 is updated using the updating rules. Subsequent tableaus are also constructed and shown below:

Table 3

Profit		C_j	70	40	0	0	0	0
Variables			q_1	q_2	S_1	S_2	S_L	S_C
Profit	Variables	Solution						
1	2	3	4	5	6	7	8	9
40	q_2	20	0	1	1.0	–2.5	0	0
70	q_1	50	1	0	0	1.25	0	0
0	S_L	10	0	0	–2.0	(1.25)	1	0
0	S_C	100	0	0	–3.0	2.5	0	1
	Z_j		70	40	40	–12.5	0	0
	$Z_j - C_j$		0	0	40	–12.5	0	0
* Profit = 40 x 20 + 70 x 50 = 4300								

Pivot is 1.25.

The solution is $q_1 = 50$, and $q_2 = 20$ corresponds to point C in the graph, which is not optimal.

Table 4

Profit		C_j	70	40	0	0	0	0
Variables			q_1	q_2	S_1	S_2	S_L	S_C
Profit	Variables	Solution						
1	2	3	4	5	6	7	8	9
70	q_2	40	0	1	–3	0	2	0
40	q_1	40	1	0	2	0	–1	0
0	S_L	8	0	0	–1.6	1	0.8	0
0	S_C	80	0	0	1	0	–2	1
	Z_j		70	40	20	0	10	0
	$Z_j - C_j$		0	0	20	0	10	0
* Profit = 70 x 40 + 40 x 40 = 4400								

The optimal solution is reached as all ($Z_j - C_j$) values are non-negative.

The optimal solution is:

$q_1 = 40$

$q_2 = 40$

$S_2 = 8$

$S_C = 80$

$S_1 = 0$

$S_L = 0$

The maximum profit is Rs.4,400, that is, produce 40 of type A castings and 40 of type B castings and we are left with 8 kgs of raw material-2 and 80 units of unused foundry capacity. Raw material-1 and labor will be completely utilized.

Example 6

Consider a firm producing batteries for cars and trucks. Each car battery costs Rs.400 in materials and machine time plus Rs.200 in wages and each truck battery costs Rs.650 in materials and machine time plus Rs.150 in wages. The selling prices of a car and a truck battery are Rs.700 and Rs.1,000 respectively. The firm is able to sell as many units as it can produce. The firm wants to plan its next months' production. The firm has 2,100 hours of machine time and 1,000 hours of assembly time available in the next month. The production of each car battery requires 10 hours of machine time and 10 hours of assembly time. The production of each truck battery requires 30 hours of machine time and 10 hours of assembly time. The firm arrived at a forecast of the cash balance of Rs.72,000, which can be used to meet the expenses of materials and wages. Assume that the payment for materials and wages has to be made in the month and the firm cannot get any more cash. With these production and financial constraints, the firm has to determine the product mix which will give maximum profit next month.

Let x and y be the number of car and truck batteries respectively, to be produced next month. The profit on a car battery is sale price minus the cost of materials and wages, that is, 700 – 400 – 200 = 100. Similarly, the profit on a truck battery is 1,000 – 650 – 150 = 200. The problem is to find the values of x and y that will maximize the profit. The formulation is:

Maximize $100x + 200y$

Subject to:

Machine capacity : $10x + 30y \leq 2{,}100$

Assembly capacity : $10x + 10y \leq 1{,}000$

Cash availability : $600x + 800y \leq 72{,}000$

$x \geq 0,\ y \geq 0$

For applying the simplex method, we convert the inequalities into equalities by adding the slack variables S_1, S_2, and S_3 for the constraints.

The problem now is:

Maximize $100x + 200y + 0S_1 + 0S_2 + 0S_3$

Subject to:

$$10x + 30y + S_1 = 2{,}100$$
$$10x + 10y + S_2 = 1{,}000$$
$$600x + 800y + S_3 = 72{,}000$$
$$x \geq 0,\ y \geq 0,\ S_1 \geq 0,\ S_2 \geq 0,\ S_3 \geq 0$$

The tablea are constructed below to obtain the optimal solution. The pivots identified in each tableau are circled.

Table 1

Profit			100	200	0	0	0
Variables			x	y	S_1	S_2	S_3
Profit	Variables	Solution					
0	S_1	2,100	10	(30)	1	0	0
0	S_2	1,000	10	10	0	1	0
0	S_3	72,000	600	800	0	0	1
	$Z_j - C_j$	0	–100	–200	0	0	0

Table 2

Profit			100	200	0	0	0
Variables			x	y	S_1	S_2	S_3
Profit	Variables	Solution					
200	y	70	$\frac{1}{3}$	1	$\frac{1}{30}$	0	0
0	S_2	300	$\frac{20}{3}$	0	$-\frac{1}{3}$	1	0
0	S_3	16,000	$\frac{1,000}{3}$	0	$-\frac{80}{3}$	0	1
	$Z_j - C_j$	14,000	$\frac{-100}{3}$	0	$\frac{20}{3}$	0	0

Table 3

Profit			100	200	0	0	0
Variables			x	y	S_1	S_2	S_3
Profit	Variables	Solution					
200	Y	55	0	1	$\frac{1}{20}$	$-\frac{1}{20}$	0
100	X	45	1	0	$\frac{1}{-20}$	$\frac{3}{20}$	0
0	S_3	1,000	0	0	–10	–50	1
	$Z_j - C_j$	15,500	0	0	5	5	0

The optimal solution is:

x = 45 and y = 55 and the maximum profit = 15,500, that is, the firm needs to produce 45 car batteries and 55 truck batteries.

Minimization

Example 7

Let us consider the minimization problem solved graphically earlier.

Minimize $Z = 60q_1 + 80q_2$

subject to constraints

$20q_1 + 30q_2 \geq 900$

$40q_1 + 30q_2 \geq 1200$

$q_1 \geq 0, q_2 \geq 0$

The solution for minimization is similar to that of maximization except that, we introduce some new variables to convert inequalities into equalities. The variable we use to convert the 'greater than' type of inequality into an equation is called 'surplus variable' and it represents the excess of what is generated over the requirement. To convert the inequality into an equality, we should subtract the surplus variable from the LHS.

Now we have,

$20q_1 + 30q_2 - S_1 = 900$

$40q_1 + 30q_2 - S_2 = 1200$

If the values of q_1 and q_2 are equal to zero, we get $S_1 = -900$ and $S_2 = -1200$. This, however, is not feasible as it violates the non-negativity restriction. To avoid this, we add artificial variables to the LHS.

These artificial variables do not represent any quantity relating to the decision problem hence, they must be removed in the last solution. If they appear in the final solution, it represents a situation of infeasibility. To ensure that this does not take place, we assign a value M, which is very large, to each artificial variable. M represents a number higher than any finite number. Hence, this method is also called the Big M method.

Now, the objective function and subject to constraints are written as

Minimize $Z = 60q_1 + 80q_2 + 0S_1 + 0S_2 + MA_1 + MA_2$

Subject to constraints

$20q_1 + 30q_2 - 1S_1 + 0S_2 + 1A_1 + 0A_2 = 900$

$40q_1 + 30q_2 + 0S_1 - 1S_2 + 0A_1 + 1A_2 = 1200$

The initial simplex table is given below. Note how we introduce artificial variables first in the initial solution.

Table 1

Cost	C_j	60	80	0	0	M	M	Minimum Ratio			
Variables		q_1	q_2	S_1	S_2	A_1	A_2				
Basic Variable	Co-efficient of Basic Variable	Solution Values									
A_1	M	900	20	30		–1	0	1	0	900/20	
A_2	M	1200	(40)	30		0	–1	0	1	1200/40	Key Row
		Z_j	60M	60M		–M	–M	M	M		
		$C_j - Z_j$	60 – 60M	80 – 60M		M	M	0	0		

Key Column

Explanation (Table 1)

1. In a minimization problem, we compute $C_j - Z_j$, instead of $Z_j - C_j$, as in the case of a maximization problem. The rest of the procedure remains the same.
2. Finding the variable to 'enter solution': This is done by identifying the key column (K.C). The column corresponding to the most negative value in the $(C_j - Z_j)$ row is called the key column in table 1.
3. Find the variable to 'leave solution': This is done by identifying the key row. As in a maximization problem, the row corresponding to the minimum positive ratio is the key row where the minimum ratio is calculated by dividing the solution values with corresponding values of the key column.
4. Optimal level is reached when the M values, if any, in the $(C_j - Z_j)$ row are positive.

Table 2

Cost		C_j	60	80	0	0	M	M	Minimum Ratio	
Variables			q_1	q_2	S_1	S_2	A_1	A_2		
Basic Variable	Co-efficient of Basic Variable	Solution Values								
A_1	M	300	0	(15)	–1	1/2	1	–1/2	20	Key Row
q_2	60	30	1	¾	0	–1/40	0	1/40	40	
		Z_j	60	15M + 45	–M	M/2 – 3/2	M	–M/2 + 3/2		
		$C_j - Z_j$	0	35 – 15M	M	3/2 – M/2	0	–3/2 + M/2		
				Key Column						

Explanation (Table 2)

1. The values in table 2 are updated in the same manner as in a maximization problem.
2. For example, Z_j value of q_2 column is computed as

   $15 \times M + 60 \times 3/4 = 15M + 45$
3. The corresponding $C_j - Z_j$ value is

   $80 - (15M + 45) = -15M + 35$
4. The variable which corresponds to the most negative $(C_j - Z_j)$ value is the one that enters the next solution. Here, the most negative $C_j - Z_j$ value is $35 - 15M$ and the corresponding variable that enters the next solution is q_2.
5. The variable that leaves the solution is the one corresponding to the least positive ratio. Here, the minimum positive ratio is 20 and the corresponding variable that leaves the solution is A_1.

Table 3

Cost		C_j	60	80	0	0	M	M	Minimum Ratio
Variables			q_1	q_2	S_1	S_2	A_1	A_2	
Basic Variable	Coefficient of Basic Variable	Solution Values							
q_2	80	20	0	1	–1/15	1/30	1/15	–1/30	
q_1	60	15	1	0	1/20	–1/20	–1/20	1/20	
		Z_j	60	80	–7/3	–1/3	7/3	1/3	
		$C_j - Z_j$	0	0	7/3	1/3	M – 7/3	M – 1/3	

As all $C_j - Z_j$ values are non-negative, we have reached the optimal solution.

The optimal solution is $q_1 = 15$ and $q_2 = 20$. The lowest cost is calculated by substituting the solution values in the objective function.

Minimize $Z = 60 \times 15 + 20 \times 80 = 2500$

The least cost is Rs.2,500.

POST OPTIMAL ANALYSIS

It can be seen from the optimal solution for the foundry problem that two resources, raw material-1 and labor, are exhausted whereas the other two resources, raw material-2 and foundry capacity, remain available. This implies that the availability of raw material-1 and labor are both exerting a restrictive effect on foundry operation and its profitability. Let us suppose, the foundry can buy the raw material-1 in the open market at a cost of Rs.15 per kg; then, is it worth buying and increasing its production to make more profit? Similarly, if the foundry can hire extra labor for Rs.15 per day, then, is it worth hiring the extra labor?

The above questions can be answered from the $(Z_j - C_j)$ values in the final tableau. The $(Z_j - C_j)$ value corresponding to variables S_1, that is, column 6 is 20. This indicates that the profit can be increased by Rs.20 for a unit increase in the availability of raw material-1. Thus, if one kilogram of raw material-1 costs Rs.15, then by purchasing it and changing the product mix, the foundry can increase its profit by Rs.20 – Rs.15 = Rs.5. Similarly, the $(Z_j - C_j)$ value corresponding to the variable S_L is 10 and this indicates that for a unit increase in the availability of labor, profit can be increased by Rs.10. Hence, it is worth hiring labor if its cost is less than Rs.10. Similarly, it is worth buying raw material-1 from open market, if its cost is less than Rs.20 per kg.

The information in the final tableau is also useful in studying the effects of the variations in the profit contributions on the product mix.

DUALITY

For every LP formulation, there exists another unique linear programming formulation called the 'Dual' (the original formulation is called the 'Primal'). The same data can be used for both 'Dual' and 'Primal' formulation. Both can be solved in a similar manner as the Dual is also an LP formulation.

The Dual can be considered as the 'inverse' of the Primal in every respect. The column coefficients in the Primal constraints become the row co-efficients in the Dual constraints. The coefficients in the Primal objective function become the right-hand-side constraints in the Dual constraints. The column of constants on the right hand side of the Primal constraints becomes the row of coefficients of the dual objective function. The direction of the inequalities are reversed. If the primal objective function is a 'Maximization' function then the dual objective function is a 'Minimization' function and vice versa.

Example 8

Consider the following 'Primal' LP formulation.

Maximize $12x_1 + 10x_2$

Subject to $2x_1 + 3x_2 \leq 18$

$2x_1 + x_2 \leq 14$

$x_1, x_2 \geq 0$

The 'Dual' formulation for this problem would be

Minimize $18y_1 + 14y_2$

Subject to $2y_1 + 2y_2 \geq 12$

$3y_1 + y_2 \geq 10$

$y_1 \geq 0, y_2 \geq 0$

Note the following:

1. The column coefficient in the Primal constraint namely (2,2) and (3,1) have become the row coefficient in the Dual constraints.
2. The coefficient of the Primal objective function namely, 12 and 10 have become the constants in the right-hand-side of the Dual constraints.
3. The constants of the Primal constraints, namely 18 and 14, have become the coefficients in the Dual objective function.
4. The direction of the inequalities have been reversed. The Primal constraints have the inequalities of $\leq$ while the Dual constraints have the inequalities of $\geq$.
5. While the Primal is a 'Maximization' problem, the Dual is a 'Minimization' problem and vice versa.

Why the Dual Formulation?

Dual formulation is done for a number of reasons. The solution to the Dual problem provides all essential information about the solution to the Primal problem. A solution for the LP problem can be determined either by solving the original problem or the Dual problem. Sometimes, it may be easier to solve the Dual problem rather than the Primal problem as the primal problem involves few variables but many constraints.

Remark

In the above sections, we have learnt to apply the simplex method to problems where the objective function is to be maximized. What if the problem were to

Minimize 50x – 70y?

This is equivalent to

$$-\{\text{Maximize} -(50x - 70y)\} = -\{\text{Maximize } -50x + 70y\}$$

We apply the Simplex Method and the final solution is obtained by taking the negative of the optimum solution (of the Maximization problem).

Comparing the Optimal Solutions of the Primal and Dual

Let us consider the example discussed under minimization. If it is considered as Primal, then the Dual is a maximization problem.

Primal	Dual
Minimize Z	Maximize $\overline{Z}$
$= 60q_1 + 80q_2$	$= 900y_1 + 1200y_2$
Subject to	Subject to
$20q_1 + 30q_2 \geq 900$	$20y_1 + 40y_2 \leq 60$
$40q_1 + 30q_2 \geq 1200$	$30y_1 + 30y_2 \leq 80$
$q_1, q_2 \geq 0$	$y_1, y_2 \geq 0$

The simplex table containing the optimal solution to the primal which was discussed earlier is as follows.

Simplex Table: Optimal Solution of the Primal

Cost		C_j	60	80	0	0	M	M
Variables			q_1	q_2	S_1	S_2	A_1	A_2
Basic Variable	Coefficient of Basic Variable	Solution Values						
q_2	80	20	0	1	–1/15	1/30	1/15	–1/30
q_1	60	15	1	0	1/20	–1/20	–1/20	1/20
		Z_j	60	80	–7/3	–1/3	7/3	1/3
		$C_j - Z_j$	0	0	7/3	1/3	M – 7/3	M – 1/3

Let us now consider the solution to the Dual problem.

Introducing slack variables in the corresponding maximization problem, we get

Maximize $\overline{Z} = 900y_1 + 1200y_2 + 0S_1 + 0S_2$

Subject to $20y_1 + 40y_2 + 1S_1 + 0S_2 = 60$

$30y_1 + 30y_2 + 0S_1 + 1S_2 = 80$

In the initial solution, we introduced the basic variables S_1 and S_2 (as shown in the following Table 1).

On comparing the optimal solutions of the Primal and Dual problems it may be noted that there is a correspondence between their variables.

In effect, the following observations are made:

a. The optimal solution objective function value is the same for both Primal and Dual. $q_1 = 15$ and $q_2 = 20$, therefore, minimize $Z = 15 \times 60 + 80 \times 20 = 2500$. Similarly $y_1 = 7/3$ and $y_2 = 1/3$, therefore, maximize $\overline{Z} = 900 \times 7/3 + 1200 \times 1/3 = 2500$.

Table 1

Profit		C_j	900	1200	0	0		
Variables			y_1	y_2	S_1	S_2		
Basic Variable	Coefficient of Basic Variable	Solution Values					Minimum Ratio	
S_1	0	60	20	(40)	1	0	3/2 = 1.5	Key Row
S_2	0	80	30	30	0	1	8/3 = 2.6	
		Z_j	0	0	0	0		
		$Z_j - C_j$	–900	–1200	0	0		
				Key Column				

Table 2

Profit		C_j	900	1200	0	0		
Variables			y_1	y_2	S_1	S_2		
Basic Variable	Coefficient of Basic Variable	Solution Values					Minimum Ratio	
y_2	1200	3/2	1/2	1	1/4 0	0	3	
S_2	0	35	15	(0)	– 3/4	1	7/3	Key Row
		Z_j	600	1200	30	0		
		$Z_j - C_j$	–300	0	30	0		
			Key Column					

Table 3: Optimal Solution of the Dual

Profit		C_j	900	1200	0	0
Variables			y_1	y_2	S_1	S_2
Basic Variable	Coefficient of Basic Variable	Solution Values				
y_2	1200	1/3	0	1	1/20	–1/30
y_1	900	7/3	1	0	–1/20	1/15
		Z_j	900	1200	15	20
		$Z_j – C_j$	0	0	15	20

b. The optimal solution values of the dual variables (7/3, 1/3) are the coefficients of the slack variables in the Δ_j row of the primal. Thus, in the Dual $y_1 = 7/3$ and $y_2 = 1/3$, whereas in the Primal $S_1 = 7/3$ and $S_2 = 1/3$ in the Δ_j row. As we have seen, $\Delta_j = Z_j – C_j$ for any maximization problem, whereas $\Delta_j = C_j – Z_j$ for any minimization problem.

c. Values in the Δ_j row under columns S_1 and S_2 of the optimum table of the Dual are the same as values corresponding to the solution variables q_1 and q_2 in the optimal table of Primal.

Hence, we can conclude that optimum solution to a primal can always provide solution to its Dual and vice-versa. Hence, both Primal and Dual need not be solved to obtain the solution. This is a computational advantage in some situations.

INTEGER PROGRAMMING

An integer-programming problem is identical to a linear programming problem except that one or more decision variables are constrained to take integer values. Such problems cannot be solved by the simplex method. They are solved by specialized procedures which are computer intensive.

GOAL PROGRAMMING

This provides a more realistic model. In a modern setting, profit maximization may not be the only objective of a business concern. Other objectives or goals could be sound ecological management, networking in the neighborhood and maximizing market share. These goals may dominate the earlier objective.

The idea is that a decision maker may not always be searching for an optimal solution but a “satisfying” solution that attempts to satisfy the many concerns of the management. Prof. Herbert A. Simon felt that a manager may not be able to optimize but may have to “satisfy”.

Simplex method requires one goal. The goals are weighted and a single objective function is constructed which is then optimized to solve a goal-programming problem. Goal programming trades-off among the goals until the most satisfying solution is found.

Application of GPP

1. For solving the problem of natural resources.
2. For solving the problem of forest production.
3. For solving the problem of agriculture.
4. For solving the planning problem of water resourcesing.

ADDITIONAL ILLUSTRATIONS

Illustration 1

For the following problem, use graphical method and find the solution.

Maximize $4x - 8y$

Subject to

$-5x + 4y \leq 10$

$5x + 3y \leq 95$

$x + 2y \geq 26$

$x \geq 0, y \geq 0$

Solution

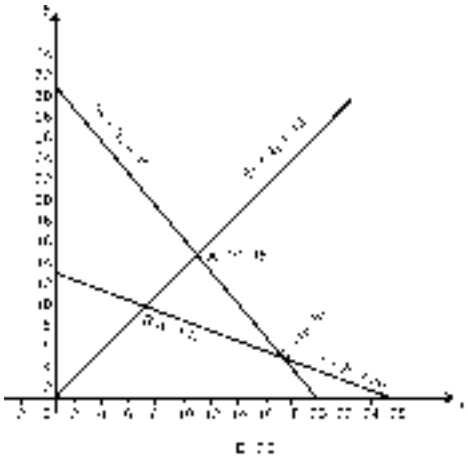


Illustration 2

ABC corporation plans to produce two products, I and H, which will be stored in a storage area whose capacity is 30,000 square feet. Product I takes 3 square feet of space per unit, and product II requires 4 square feet. It takes four machine hours to manufacture a unit of product I, while 8 machine hours are required for a unit of product II. A total of 48,000 machine hours is available. Also available are 36,000 man hours for finishing the products. One unit of product I takes 4 man hours for finishing, while one unit of product II needs 3 man hours.

Profit contribution is Rs.20 per unit of product I and Rs.60 per unit of Product II. In order to maximize profit for the period, how many units of each product should ABC corporation manufacture? Solve the problem graphically with the help of the simplex method.

Solution

Let X_1 – Number of units of product I to be manufactured.

X_2 – Number of units of product II to be manufactured.

The problem is formulated as shown below:

Maximize $Z = 20 X_1 + 60 X_2$

Subject to

$3 X_1 + 4 X_2 \leq 30{,}000$ (Storage space)

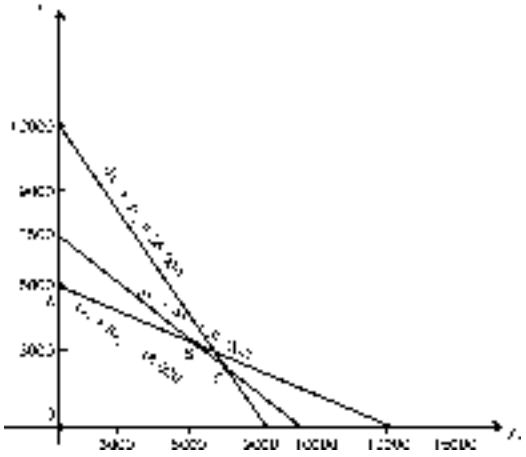
$4 X_1 + 8 X_2 \leq 48{,}000$ (Machine hours)

$4 X_1 + 3 X_2 \leq 36{,}000$ (Man hours)

$X_1, X_2 \geq 0$

We have to represent these inequalities graphically. This can be done by replacing the $\leq$ sign by = sign. The point of intersection of the line $3X_1 + 4X_2 = 30,000$ with X_1 and X_2 axis.

The point of intersection of $3X_1 + 4X_2 = 30,000$ and X_1 axis is got by substituting $X_2 = 0$ (Equation of X1 axis). The point is (10,000, 0). The point of intersection of $3X_1 + 4X_2 = 30,000$ and X_2 axis is got by substituting $X_1 = 0$ (Equation of X_2 axis). The point is (0, 7,500). We repeat this process for other inequalities and draw the lines. The points of intersection of the line $4X_1 + 8X_2 = 48,000$ and X_1 axis is (12000, 0) and with X_2 axis is (0, 6000).



The points of intersection of $4X_1 + 3X_2 = 36,000$ with X_1 and X_2 axis are (9,000, 0) and (0, 12000).

The feasible region is given by the corner points O = (0,0); D = (9000, 0); C $= \left[\frac{54000}{7}, \frac{12000}{7}\right]$, B = (6000, 3000), and A = (0, 6000)

At A, Z = (20)(0) + (60)(6000) = 36,0000

At B, Z = (20)(6000) + (60)(3000) = 300000

At C, Z = (20)(7714.285) + 60 (1714.285) = 257142.8

At D, Z = (20)(9000) + (0)(60) = 180000

At O, Z = 0

The maximum profit of Rs.3,60,000 is attained at the point (0, 6000).

Simplex Method:

Introduce the slack variables to convert the inequalities into equations.

The LPP problem is restated as

Maximize Z =

$$20X_1 + 60X_2 + 0S_1 + 0S_2$$

Subject to

$$3X_1 + 4X_2 + S_1 = 30,000$$

$$4X_1 + 8X_2 + S_2 = 48,000$$

$$4X_1 + 3X_2 + S_3 = 36,000$$

$$X_1, X_2 \geq 0.$$

Table 1

		Profit	20	60	0	0	0
		Variables	X_1	X_2	S_1	S_2	S_3
Profit	Variables	Solution					
0	S_1	30,000	3	4	1	0	0
0	S_2	48,000	4	8	0	1	0
0	S_3	36,000	4	3	0	0	1
	$Z_j - C_j$	0	–20	–60	0	0	0

By observing $Z_j - C_j$ row, we note that the profits can be increased by increasing the production level as by not producing anything we are losing Rs.20 and Rs.60 per unit respectively on product X_1 and X_2.

We begin with product X_2 since the profit can be increased at the rate of Rs.60 per unit. Product X_2 becomes the basic variable now. The production of X_2 can be increased until one of the resources gets exhausted. The minimum positive ratio of elements of column (3) and (2) (that is of product X_2) will indicate the resource which will be first exhausted.

From,

$$\frac{30,000}{4} = 7,500; \frac{48,000}{8} = 6,000; \text{ and } \frac{36,000}{3} = 12,000,$$

We note that the resource, machine hours, gets exhausted first by producing 6000 units of product X_2.

As S_2 = 0, now, we replace it with X_2. The pivot element is the element common to the column X_2 and row X_2. In this case, it is 8, and we circle it. Now, we are ready to update the table. This is done in the following steps: (i) Fill in row 1 and row 2 of table 2 in the same way as table 1. Fill in the column (2) replacing the slack variables selected with X_2, (ii) Fill in column (1) with the profit contribution of the variables in column (2), (iii) Update the pivot row by dividing each element by the pivot, (iv) Update the pivot column by filling it with zero's. The other elements are updated as follows: the first element 30000 in column 3 is replaced by the number obtained from the following operations.

We have, 30000 4

(1) Pivot element (4)

From 30000 we substract the product 4 and 6000. Respective signs should be considered while computing this,

= 30000 – (4)(6000)

= 30000 – 24000

= 6000

Other elements are updated similarly. We repeat this until the elements in the $Z_j - C_j$ row are all positive. This indicates that an optimal solution has been reached.

		Profit	20	60	0	0	0
		Variables	X_1	X_2	S_1	S_2	S_3
Profit	Variables	Solution					
0	S_1	6,000	1	0	1	–1/2	0
60	X_2	6,000	1/2	1	0	1/8	0
0	S_3	18,000	5/2	0	0	–3/8	1
	$Z_j - C_j$	360000	10	0	0	15/2	0

The profit is maximized when 6000 units of X_2 are manufactured. This is same as one obtained from the graphical method.

Illustration 3

Maximize Z

where

$$Z = 3X_1 + 2X_2$$

Subject to

$$X_1 + X_2 \leq 15$$
$$2X_1 + X_2 \leq 28$$
$$X_1 + 2X_2 \leq 20$$
$$X_1 \geq, X_2 \geq 0$$

Solve the above LPP problem by Simplex Method.

Solution

Introducing the slack variables the LPP can be restated as follows:

$$\text{Maximize } Z = 3X_1 + 2X_2 + 0S_1 + 0S_2$$

Subject to

$$X_1 + X_2 + S_1 = 15$$
$$2X_1 + X_2 + S_2 = 28$$
$$X_1 + 2X_2 + S_3 = 20$$
$$X_1, X_2 \geq 0.$$

Simplex Table

		Profit	3	2	0	0	0
		Variables	X_1	X_2	S_1	S_2	S_3
Profit	Variables	Solution					
0	S_1	15	1	1	1	0	0
0	S_2	28	**2**	1	0	1	0
0	S_3	20	1	2	0	0	1
	$Z_j - C_j$	0	–3	–2	0	0	0
0	S_1	1	0	½	1	–	0
3	X_1	14	1	**1/2**	0	1/2	0
0	S_3	6	0	3/2	0	1/2	1
						–	
						1/2	
	$Z_j - C_j$	42	0	–1/2	0	3/2	0
2	X_1	13	1	0	–1	1	0
3	X_2	2	0	1	2	–1	0
0	S_3	3	0	0	–3	1	1
	$Z_j - C_j$	43	0	0	1	1	0

The profit is maximized at $X_1 = 13$ and $X_2 = 2$.

Illustration 4

A manufacturer has three machines A, B and C, which are used to make three products I, II and III. A unit of product I requires 60 minutes on machine A, 60 minutes on machine B, and 40 minutes on machine C. A unit of product II requires 40 minutes on machine A, 70 minutes on machine B and 50 minutes on machine

C. A unit of product III requires 10 minutes on machine A, 30 minutes on machine B and 120 minutes on machine C. The profit contribution per unit of product I, II and III respectively are Rs.400, Rs.300 and Rs.500. The number of minutes available on each machine during a given production period are as follows:

Machine	No. of Minutes Available
A	2000
B	1000
C	1500

Using simplex method, find out the number of units of each product to be manufactured so as to maximize total profits

Solution

The LPP is formulated as shown below:

Maximize Z = $40X_1 + 30X_2 + 50X_3$

Subject $6X_1 + 4X_2 + X_3 \le 200$

$6X_1 + 7X_2 + 3X_3 \le 100$

$4X_1 + 5X_2 + 12X_3 \le 150$

$X_1, X_2, X_3 \ge 0$

[After dividing the values in three inequalities by 10]

Introducing the slack variables, the LPP can be restated as follows:

Maximize $Z = 40X_1 + 30X_2 + 50X_3 + 0S_1 + 0S_2 + 0S_3$

Subject to

$6X_1 + 4X_2 + X_3 + S_1 = 200$

$6X_1 + 7X_2 + 3X_3 + S_2 = 100$

$4X_1 + 5X_2 + 12X_3 + S_3 = 150$

$X_1, X_2, X_3 \ge 0.$

Simplex Table

		Profit	40	30	50	0	0	0
		Variables	X_1	X_2	X_3	S_1	S_2	S_3
Profit	Variables	Solution						
0	S_1	200	6	4	1	1	0	0
0	S_2	100	6	7	3	0	1	0
0	S_3	150	4	5	**12**	0	0	1
	$Z_j - C_j$	0	–40	–30	–50	0	0	0
0	S_1	375/2	17/3	43/12	0	1	0	–1/12
3	S_2	125/2	5	**23/4**	0	0	1	–1/4
0	X_3	50/4	1/3	5/12	5/12	0	0	1/12

		Profit	40	30	50	0	0	0
		Variables	X_1	X_2	X_3	S_1	S_2	S_3
	$Z_j - C_j$	625	–70/3	–110/12	0	0	0	50/12
2	S_1	10250/69	**176/69**	0	0	1	-43/69	5/69
3	X_2	250/23	20/23	1	0	0	4/23	–1/23
0	X_3	550/69	–2/69	0	1	0	–5/69	7/69
	$Z_j - C_j$	50000/69	–1060/69	0	0	1	110/69	260/69
0	S_1	8050/69	0	44/15	0	1	–	1/5
40	X_1	12.5	1	**23/20**	0	0	391/345	–1/20
50	X_3	575/69	0	1/30	1	0	1/5	1/10
	$Z_j - C_j$	63250/69	0	53/6	0	0	–1/15 322/69	3

The maximum profit is Rs.9,166.67 which is attained by manufacturing 13 units of X_1 and 8 units of X_3.

Illustration 5

Company has 24,000 kgs of chemical X and 60,000 kgs of chemical Y available per month. The company manufactures 3 industrial compounds A, B and C for unit weight of the compounds produced chemicals X and Y are required in the following proportions.

Compound A 20% of X, and 80% of Y

Compound B 40% of X, and 60% of Y

Compound C 37.5% of X, and 62.5% of Y

The company makes profit of Rs.8, Rs.6 and Rs.10 on every kg of the compound A, B and C respectively. Find the quantities of A, B and C to be manufactured in order to maximize the total profits.

Solution

Let X_1 – Quantity of A to be manufactured

X_2 – Quantity of B to be manufactured

X_3 – Quantity of C to be manufactured.

The problem can be formulated as follows:

Maximize $Z = 8X_1 + 6X_2 + 10X_3$

Subject $X_1 + 2X_2 + 3X_3 \le 24$

$4X_1 + 3X_2 + 5X_3 \le 60$

$X_1, X_2, X_3 \ge 0$

we now introduce the slack variable. The LPP can be restated as:

Maximize $Z = 8X_1 + 6X_2 + 10X_3 + 0S_1 + 0S_2$

Subject to

$X_1 + 2X_2 + 3X_3 + S_1 = 24$

$4X_1 + 3X_2 + 5X_3 + S_2 = 60$

$X_1, X_2, X_3 \ge 0.$

Simplex Table

		Profit	8	6	10	0	0
		Variables	X_1	X_2	X_3	S_1	S_2
Profit	Variables	Solution					
0	S_1	24	1	2	3	1	0
0	S_2	60	4	3	5	0	1
	$Z_j - C_j$	0	-8	-6	-10	0	0
10	X_3	8	1/3	2/3	1	1/3	0
0	S_2	20	7/3	$-1/3$	0	$-5/3$	1
	$Z_j - C_j$	80	$-14/3$	2/3	0	10/3	0
10	X_3	36/7	0	5/7	1	4/7	$-1/7$
8	X_1	60/7	1	$-1/7$	0	$-5/7$	3/7
	$Z_j - C_j$	120	0	0	0	0	2

Illustration 6

Write the dual of the following problem and solve it by simplex method.

Minimize $P = 12\,W_1 + 20\,W_2$

Subject to $W_1 + W_2 \geq 2$

$3\,W_1 + W_2 \geq 6$

$2\,W_1 + W_2 \geq 4,\ W_1, W_2 \geq 0.$

Solution

The dual of this problem would be

Maximize $Z = 2X_1 + 6X_2 + 4X_3$

Subject to

$X_1 + 3X_2 + 2X_3 \leq 12$

$X_1 + X_2 + X_3 \leq 20$

$X_1, X_2, X_3 \geq 0$

When the slack variables are introduced, the LPP can be restated as follows:

Maximize $Z = 2X_1 + 6X_2 + 14X_3 + 0S_1 + 0S_2$

Subject to

$X_1 + 3X_2 + 2X_3 + S_1 = 12$

$X_1 + X_2 + X_3 + S_2 = 20$

$X_1, X_2, X_3 \geq 0.$

Simplex Table

		Profit	2	6	4	0	0
		Variables	X_1	X_2	X_3	S_1	S_2
Profit	Variables	Solution					
0	S_1	12	1	**3**	2	1	0
0	S_2	20	1	1	1	0	1
	$Z_j - C_j$	0	-2	-6	-4	0	0
6	X_2	4	1/3	1	2/3	1/3	0
0	S_2	16	2/3	0	1/3	$-1/3$	1
	$Z_j - C_j$	24	0	2/3	0	2	0

Illustration 7

A television company has three major departments for manufacture of its two models A and B. Monthly capacities are as follows:

Department	Time in hours per unit of		Hours available in this month
	Model A	Model B	
I	4.0	2.0	1600
II	2.5	1.0	1200
III	4.5	1.5	1600

The marginal profit on one unit of model A is Rs.400 and that of model B is Rs.100. Assuming that the company can sell any quantity of either product due to favorable market conditions, determine the optimum output for both the models, the highest possible profit for this month and the slack time in the three departments.

Solution

Let X_1 and X_2 be the number of TV sets of models A and B that can be manufactured. The given problem when converted into can LPP will read as follows:

Maximize $Z = 400X_1 + 100X_2$

Subject to

$4X_1 + 2X_2 \le 1600$

or $2X_1 + X_2 \le 800$ … (1)

$2.5X_1 + X_2 \le 1200$

or $5X_1 + 2X_2 \le 2400$ … (2)

$4.5X_1 + 1.5X_2 \le 1600$

or $9X_1 + 3X_2 \le 3200$ … (3)

$X_1, X_2 \ge 0$

Introducing slack variable S_1, S_2 and S_3, the LPP can be restated as follows:

Maximize $Z = 400X_1 + 100X_2 + 0.S_1 + 0.S_2 + 0.S_3$

Subject to

$2X_1 + X_2 + S_1 = 800$

$5X_1 + 2X_2 + S_2 = 2,400$

$9X_1 + 3X_2 + S_3 = 3,200$

$X_1, X_2, S_1, S_2, S_3 \ge 0$

Simplex Table

		Profit	400	100	0	0	0
		Variables	X_1	X_2	S_1	S_2	S_3
Profit	Variables	Solution					
0	S_1	800	2	1	1	0	0
0	S_2	2400	5	2	0	1	0
0	S_3	3200	**9**	3	0	0	1
	$Z_j - C_j$	0	– 400	– 100	0	0	0
0	S_1	800/9	0	1/3	1	0	– 2/9
0	S_2	5600/9	0	1/3	0	1	– 5/9
400	X_1	3200/9	1	1/3	0	0	1/9
	$Z_j - C_j$	142222.2	0	100/3	0	0	400/9

Since all $Z_j - C_j$ are non-negative, the optimal solution is attained. The optimal values of the variables are X_1 = 355.6 units and X_2 = 0.

Therefore, the television company needs to manufacture 355.6 units of model A to achieve the maximum profit of Rs.1,42,222.20. at this level of output, unutilized capacities exist in Department I and II to the extent of 177.8 hours and 311.21 hours respectively.

Illustration 8

Obtain the dual of the following linear programming problem.

Minimize $Z = 5x_1 - 6x_2 + 4x_3$

Subject to the constraints:

$$3X_1 + 4X_2 + 6X_3 \geq 9$$

$$X_1 + 3X_2 + 2X_3 \geq 5$$

$$7X_1 - 2X_2 - X_3 \leq 10$$

$$X_1 - 2X_2 + 4X_3 \geq 4$$

$$2X_1 + 5X_2 - 3X_3 \geq 3$$

$$X_1, X_2, X_3 \geq 0.$$

Solution

In this problem, one of the primal constraints (namely $7X_1 + 2X_2 - X_3 \leq 10$) is a "$\leq$"constraint while all the others are"$\geq$" constraints. The dual cannot be worked out unless all the constraints are in the same direction. To convert this into"$\geq$"constraint, multiply both the sides of the equation by "– sign,. After multiplying the constraint by "–" sign, it will become – $7X_1 + 2X_2 + X_3 \geq -10$. Now, all the constraints are in the same direction and the dual can be worked out.

The dual formulation is Maximise $Z = 9w_1 + 5w_2 - 10w_3 + 4w_4 + 3w_5$

Subject to the constraints:

$$3w_1 + w_2 - 7w_3 + w_4 + 2w_5 \leq 5$$

$$4w_1 + 3w_2 + 2w_3 - 2w_4 + 5w_5 \leq -6$$

$$6w_1 + 2w_2 + w_3 + 4w_4 - 3w_5 \leq 4$$

$$w_1, w_2, w_3 w_4, w_5 \geq 0.$$

Illustration 9

A manufacturer of leather belts makes three types of belts A, B and C which are processed on three machines M1, M2 and M3. Belt A requires 2 hours on M1 and 3 hours on M3. Belt B requires 3 hours on M1 and 2 hours on M2 and 3 hours on M3. Belt C requires 5 hours on M2, and 4 hours on M3. The availabilities are 8 hours on M1, 10 hours on M2, and 15 hours on M3 per day. The profit gained by selling each of the belts A, B and C are Rs.3, 5 and 4 respectively. Formulate the problem as an LPP and find the optimum production of each type of belt per day so that the profit is maximum.

Solution

Let X_1, X_2 and X_3 be the units of type A, B and C belts manufactured. The given problem when converted into an LPP, will read as follows:

Maximize $Z = 3X_1 + 5X_2 + 4X_3$

Subject

$$2X_1 + 3X_2 \leq 8$$

$$2X_2 + 5X_3 \leq 10$$

$$3X_1 + 2X_2 + 4X_3 \leq 15$$

$$X_1, X_2, X_3 \geq 0$$

By introducing slack variables, the problem can be rewritten as

$$\text{Maximize } Z = 3X_1 + 5X_2 + 4X_3 + 0S_1 + 0S_2 + 0S_3$$

Subject to

$$2X_1 + 3X_2 + S_1 = 8$$

$$2X_2 + 5X_3 + S_2 = 10$$

$$3X_1 + 2X_2 + 4X_3 + S_3 = 15$$

$$X_1, X_2, X_3, S_1, S_2, S_3 \geq 0.$$

Simplex Table

		Profit	3	5	4	0	0	0	Mini.ratio
		Variables	X_1	X_2	X_3	S_1	S_2	S_3	
Profit	Variables	Solution							
0	S_1	8	2	3	0	1	0	0	8/3
0	S_2	10	0	2	5	0	1	0	10/2
0	S_3	15	3	2	4	0	0	1	15/3
	$Z_j - C_j$	0	–3	– 5	– 4	0	0	0	0
5	X_2	8/3	2/3	1	0	1/3	0	0	–
0	S_2	14/3	–4/3	0	**5**	– 2/3	1	0	14/15
0	S_3	29/3	5/3	0	4	– 2/3	0	1	29/12
	$Z_j - C_j$		– 1/3	0	4	– 5/3	0	0	
5	X_2	8/3	**2/3**	1	0	1/3	0	0	8/2
4	X_3	14/15	– 4/15	0	1	– 2/15	1/5	0	–
0	S_3	89/15	**41/15**	0	0	– 2/15	– 4/5	1	89/41
	$Z_j - C_j$		11/15	0	0	17/15	4/5		0
5	X_1	89/41	1	0	0	– 2/41	–12/41		15/41
4	X_2	50/41	0	1	0	15/41	8/41		– 10/41
3	X_3	62/41	0	0	1	– 6/41	5/41		4/41
	$Z_j - C_j$		0	0	0	45/41	24/41		11/41

Now, since all $Z_j - C_j$ values are positive, the optimum solution is reached and is given by

$X_1 = 89/41$, $X_2 = 50/41$, and $X_3 = 62/41$

The maximum profit is $Z^* = 3 \times 89/41 + 5 \times 50/41 + 4 \times 62/41 = 765/41$.

Illustration 10

A chemical firm produces automobile cleaner X and polisher Y and realizes a profit of Rs.10 and Rs.30 on each batch of X and Y respectively. Both products require processing through the same machine A and B. X requires 4 hours in machine A and 8 hours in machine B. Y requires 6 hours in machine A and 4 hours in machine B. During the forthcoming week, machine A and B have 12 hours and 16 hours of available capacity respectively. Assuming that the demand exists for both the products, how many batches of each should be produced to realize the optimal profit?

Formulate this problem as a linear programming problem and find the optimal solution by the graphical method and also by the simplex method.

Solution

The linear programming problem is

Maximize $Z = 10X_1 + 30X_2$

Subject

$4X_1 + 6X_2 \le 12$

$8X_1 + 4X_2 \le 16$

$X_1, X_2 \ge 0$

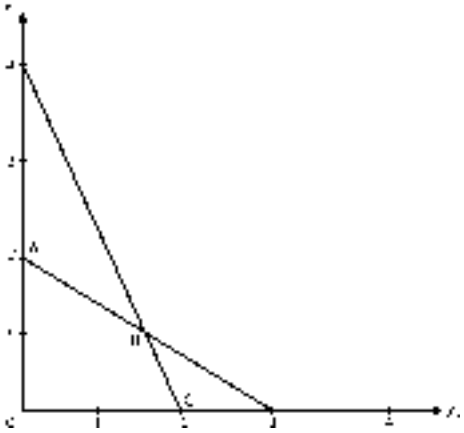
Maximize $Z = 10X_1 + 30X_2$

Subject to

$2X_1 + 3X_2 \le 6$

$2X_1 + X_2 \le 16$

$X_1, X_2 \ge 0$



At O (0,0) Z = 10(0) + 30(0) = 0

At A (0,2) Z = 10(0) + 30(2) = 60

At B (1.5,1) Z = 10(1.5) + 30(1) = 45

At C (2,0) Z = 10(2) + 30(0) = 20

Z is maximized at $X_1 = 0$, and $X_2 = 2$ equals 60.

Simplex Table

		Profit	10	30	0	0
		Variables	X_1	X_2	S_1	S_2
Profit	Variables	Solution				
0	S_1	6	2	**3**	1	0
0	S_2	4	2	1	0	1
	$Z_j - C_j$	0	– 10	– 30	0	0
30	S_1	2	2/3	1	1/3	0
0	S_2	2	4/3	0	– 1/3	1
	$Z_j - C_j$	60	10	0	10	0

Since all $Z_j - C_j$ are positive, table 2 is optimal table. The optimal values of the decision variable $X_1 = 10$, and $X_2 = 0$.

SUMMARY

- Linear programming is one of the most important techniques of operations research that is applied to a wide range of business problems. It is used in solving decision problems involving maximization of a linear objective function, subject to a set of linear constraints. Linear programming is helpful in solving various problems in finance, budgeting and investments, such as selection of a product mix which maximizes profits, determination of the capital budget which maximizes the net present value, etc.
- The Graphical method of solving linear programming problems can be applied to those problems which have only two basic variables. It is the simplest method that helps in a better understanding of the other advanced methods. But in large size linear programming problems, the solution cannot be obtained by the graphical method and a more systematic method, i.e., the Simplex method is used. The simplex method is an iterative procedure for moving from an extreme point with a low profit value to another with a higher profit value until the maximum value of the objective function is achieved. The chapter also deals with other aspects of linear programming like dual and primal formulations, post optimality analysis, integer and goal programming.

Bibliography

1. Arora, P. N., and Arora, S. *CA Foundation Course Statistics*. 6th e.d. New Delhi: S Chand & Co. Ltd., 2004.
2. Chandan, J. S. *Statistics for Business and Economics*, New Delhi: Vikas Publishing House, 1998.
3. Groebner David F., and Patrick W. Shannon. *Essentials of Business Statistics: A Decision Making Approach*. 2nd ed. New York: MacMillian College Publishing Co., 1994.
4. Gun, A. M. Gupta, M. K., and DasGupta, B. *Fundamentals of Statistics.* 7th ed. Vol. 2. Kolkata: World Press Pvt. Ltd., 2002.
5. Gupta, S. P. *Statistical Methods.* 31st Revised ed. New Delhi: Sultan Chand & Sons. 2004.
6. Hoel Paul G., and Raymond J. Jessen. *Basic Statistics for Business and Economics*. 3rd ed. New York: John Wiley & Sons, 1982.
7. Richard I. Levin, and David S. Rubin. *Statistics for Management.* 5th ed. New Delhi: Prentice Hall India Pvt. Ltd., 1994.
8. Sancheti, D. C. and Kapoor, V. K. *Statistics – Theory, Methods and Applications.* 7th ed. New Delhi: Sultan Chand & Sons, 1991.
9. Srivastava, U. K., Shenoy, G. V., and Sharma, S. C. *Quantitative Techniques for Managerial Decisions.* New Delhi: New Age International (P) Ltd., 3rd Reprint, 1995.
10. Verma, A. P. *Business Mathematics and Statistics*. New York: Asian Books Pvt. Ltd., 2002.

Glossary

A Priori Probability : Probability estimate made prior to receiving new information.

Alpha (α) : The probability of a Type I error.

Alternative Hypothesis : The conclusion we accept when the data fail to support the null hypothesis.

Analysis of Variance for Regression : The procedure for computing the F ratio used to test the significance of the regression as a whole.

Assignable Variation : It is the systematic variation present in the process. The whole system need not be redesigned in order to correct this variation.

Attributes : Qualitative variables with only two categories.

Basic Variables : Variables in the solution in a tableau. The other variables, which are out of the solution and have the value zero, are called non-basic.

Bays' Theorem : The formula for conditional probability under statistical dependence.

Bernoulli Process : A process in which each trial has only two possible outcomes, the probability of the outcome of any trial remains fixed over time, and the trials are statistically independent.

Beta (β) : The probability of a Type II error.

Bimodal Distribution : A distribution of data points in which two values occur more frequently than the rest of the values in the data set.

Binomial Distribution : A discrete distribution describing the results of an experiment known as a Bernoulli process.

Census : The measurement or examination of every element in the population.

Central Limit Theorem : A result assuring that the sampling distribution of the mean approaches normality as the sample size increases, regardless of the shape of the population distribution from which the sample is selected.

Certainty : The decision environment in which only one state of nature exists.

Cj – Zj Row : The row containing the net benefit or loss occasioned by bringing one unit of a variable into the solution of a linear programming problem.

Classical Probability : The number of outcomes favorable to the occurrence of an event divided by the total number of possible outcomes.

Clusters : Within a population, groups that are essentially similar to each other, although the groups themselves have wide internal variation.

Cluster Sampling : A method of random sampling in which the population is divided into groups, or clusters of elements, and then a random sample of these clusters is selected.

Coefficient of Correlation : The square root of the coefficient of determination. Its sign indicates the direction of the relationship between two variables, direct or inverse.

Coefficient of Determination : A measure of the proportion of variation in Y, the dependent variable, that is explained by the regression line, that is, by Y's relationship with the independent variable.

Coefficient of Multiple Correlation, R : The positive square root of R^2.

Coefficient of Multiple Determination, R^2 : The fraction of the variation of the dependent variable that ıs explained by the regression, R^2 measures how well the multiple regression fits the data.

Coefficient of Variation : A relative measure of dispersion, comparable across distributions, that expresses the standard deviation as a percentage of the mean.

Collectively Exhaustive Events : The list of events that represents all the possible outcomes of an experiment.

Combinations : Each of the groups or selections which can be made by taking some or all of a number of things is called a combination.

Common (Random) Variation : Variability inherent in a process. It cannot be reduced without redesigning the entire process.

Conditional Probability : The probability of one event occurring, given that another event has occurred.

Conditional Profit : The profit that would result from a given combination of decision alternative and state of nature.

Continuous Probability Distribution : A probability distribution in which the variable is allowed to take on any value within a given range.

Continuous Data : Data that may progress from one class to the next without a break and may be expressed by either whole numbers or fractions.

Continuous Random Variable : A random variable to take on any value within a given range.

Correlation Analysis : A technique to determine the degree to which variables are linearly related.

Cumulative Frequency Distribution : A tabular display of data showing how many observations lie above, or below, certain values.

Curvilinear Relationship : An association between two variables that is described by a curved line.

Cyclical Fluctuation : A type of variation in a time series, in which the value of the variable fluctuates above and below a secular trend line.

Data : A collection of any number of related observations on one or more variables.

Dependent Samples : Samples drawn from two populations in such a way that the elements in one sample are matched or paired with the elements in the other sample, in order to allow a more precise analysis by controlling for extraneous factors.

Dependent Variable : The variable we are trying to predict in regression analysis.

Deseasonalization : A statistical process used to remove the effects of seasonality from a time series.

Direct Relationship : A relationship between two variables in which, as the independent variable's value increases, so does the value of the dependent variable.

Discrete Data : Data that do not progress from one class to the next without a break; that is where classes represent distinct categories or counts and may be represented by whole numbers.

Discrete Probability Distribution : A probability distribution in which the variable is allowed to take on only a limited number of values.

Discrete Random Variable : A random variable that is allowed to take on only a limited number of values.

Dispersion : The spread or variability in a set of data.

Dual : A linear programming problem which is one part of a pair of associated linear programming problems called the primal and the dual.

Equation : It is a statement which shows that the two algebraical expressions considered are equal.

Estimate : A specific observed value of an estimator.

Estimating Equation : A mathematical formula that relates the dependent variable to the independent variables in regression analysis.

Expected Value : A weighted average of the outcomes of an experiment.

Expected Value of a Random Variable : The sum of the products of each value of the random variable with that value's probability of occurrence.

Expected Value of Perfect Information : The difference between expected profit (under conditions of risk) and expected profit with perfect information.

Expected-Value Criterion : A criterion requiring the decision maker to calculate the expected value for each decision alternative.

Experiment : The activity that results in, or produces, an event.

Extreme Point : A corner of the feasible region.

Feasible Region : That area containing those solutions which satisfy all the constraints in the problem.

Finite Population : A population having a limited size.

Finite Population Multiplier : A factor used to correct the standard error of the mean for studying a population of finite size that is small in relation to the size of the sample.

Fixed-Weight Aggregates Method : To weigh an aggregates index, this method uses quantities consumed during some representative period, as weights.

Fractile : In a frequency distribution, the location of a value at, or above, a given fraction of the data.

Frequency Curve : A frequency polygon smoothened by adding classes and data points to a data set.

Frequency Distribution : An organized display of data that shows the number of observations from the data set, that falls into each set of mutually exclusive and collectively exhaustive classes.

Frequency Polygon : A line graph connecting the midpoints of each class in a data set, plotted at a height corresponding to the frequency of the class.

Geometric Mean : A measure of central tendency used to measure the average rate of growth, computed by taking the 'n'th root of the product of n values representing change.

Highest Common Factor : The highest common factor of two or more algebraical expressions is the expression of highest dimensions which divides each of them without remainder.

Histogram : A graph of a data set, composed of a series of rectangles, each proportional in width to the range of values in a class and proportional in height to the fraction of items in the class.

Hypothesis : An assumption or speculation we make about a population parameter.

Identity : If two expressions are equal for any values we give to the symbols is said to be an identity.

Independent Variables : The known variable, or variables, in regression analysis.

Index Number : A ratio that measures how much a variable changes over time.

Inequality : A mathematical expression indicating that minimum or maximum requirements must be met.

Infeasibility : The condition when there is no solution which satisfies all the constraints in a problem.

Infinite Population : A population in which it is theoretically impossible to observe all the elements.

Interfractile Range : The difference between the values of two fractiles.

Interquartile Range : The difference between the values of the first and the third quartiles.

Interval Estimate : A range of values used to estimate an unknown population parameter.

Inverse Relationship : A relationship between two variables in which, as the independent variable increases, the dependent variable decreases.

Irregular Variation : A condition in a time series in which the value of a variable is completely unpredictable.

Joint Probability : The probability of two events occurring together or in succession.

Judgment Sampling : A method of selecting a sample from a population in which personal expertise is used to identify those items from the population that are to be included in the sample.

Laspeyres Method : To weigh an aggregates index, this method uses the quantities consumed during the base period, as weights.

Least-Squares Method : A technique for fitting a straight line through a set of points in such a way that the sum of the squared vertical distances from the points to the line is minimized.

Linear Programming : A mathematical technique for finding the best uses of an organization's resources.

Linear Relationship : A particular type of association between two variables that can be described mathematically by a straight line.

Mean : The arithmetic average of a set of observations.

Measure of Central Tendency : A measure indicating the value to be expected of a typical or middle data point.

Measure of Dispersion : A measure describing how the observations in a data set are scattered.

Median Class : The class in a frequency distribution that contains the median values for a data set.

Median : The middle point of a data set, a measure of location that divides the data set into halves.

Minimum Probability : The probability of selling at least an additional unit that must exist to justify stocking that unit.

Mode : The value most often repeated in the data set.

Multicollinearity : A statistical problem sometimes present in multiple-regression analysis in which the reliability of the regression coefficients is reduced, owing to a high level of correlation between the independent variables.

Multiple Regression : A statistical process by which several variables are used to predict another variable.

Mutually Exclusive Events : Events that cannot happen together.

Node : The point at which a chance event or a decision takes place on a decision tree.

Non-negativity Constraints : Constraints that restrict all the variables to be zero or positive.

Normal Distribution : A distribution of a continuous random variable with a single-peaked, bell-shaped curve. The mean lies at the center of the distribution, and the curve is symmetrical around a vertical line erected at the mean. The two tails extend indefinitely, never touching the horizontal axis.

Objective : A goal of the organization.

Objective Function : An expression which shows the relationship between the variables in the problem and the firm's goal.

Obsolescence Loss : The loss occasioned by stocking too many units and having to dispose of unsold units.

Ogive : A graph of a cumulative frequency distribution.

Paasche Method : In weighting an aggregates index, the Paasche method uses, as weights, the quantities consumed during the current period.

Parameters : Values that describe the characteristics of a population.

Percentage Relative : Ratio of a current value to a base value with the result multiplied by 100.

Permutations : Each of the arrangements which can be made by taking some or all number of things is called a permutation.

Pivoting : The process of going from one simplex tableau to the next.

Population : A collection of all the elements we are studying and about which we are trying to draw conclusions.

Posterior Probability : A probability that has been revised after additional information was obtained.

Price Index : Compares levels of prices from one period to another.

Primal : A linear programming problem which is one part of a pair of associated linear programming problems called the primal and the dual.

Probability Distribution : A list of the outcomes of an experiment with the probabilities we would expect to see associated with these outcomes.

Probability : The chance that something will happen.

Probability Tree : A graphical representation showing the possible outcomes of a series of experiments and their respective probabilities.

p Chart : A control chart used for monitoring the proportion of items in a batch that meet certain specified requirements.

Quantity Index : A measure of how much the number or quantity of a variable changes over time.

Quartiles : Fractiles that divide the data into four equal parts.

Quadratic Equation : An equation which contains the square of the unknown quantity but no higher power is called a quadratic equation.

Random or Probability Sampling : A method of selecting a sample from a population in which all the items in the population have an equal chance of being chosen in the sample.

Random Variable : A variable that takes on different numerical values as a result of the outcomes of a random experiment.

Range : The distance between the highest and lowest values in a data set.

Ratio to Moving Average Method : A method employed in time series analysis to identify the component of the seasonal variation.

Raw Data : Information before it is arranged or analyzed by statistical methods.

Reduced Costs : The entries of the Cj – Zj row.

Redundant Constraint : A constraint that has no effect on the feasible space.

Regression Line : A line fitted to a set of data points to estimate the relationship between two variables.

Regression : The general process of predicting one variable from another by statistical means, using previous data.

Relative Frequency Distribution : The display of a data set that shows the fraction or percentage of the total data set that falls into each set of mutually exclusive and collectively exhaustive classes.

Relative Frequency of Occurrence : The proportion of times that an event occurs in the long run when conditions are stable, or the observed relative frequency of an event in a very large number of trials.

Representative Sample : A sample that contains the relevant characteristics of the population in the same proportions as they are included in that population.

Sample : A portion of the elements in a population chosen for direct examination or measurement.

Sample Space : The set of all possible outcomes of an experiment.

Sampling Distribution of a Statistic : For a given population, a probability distribution of all the possible values a statistic may take on for a given sample size.

Sampling Distribution of the Mean : A probability distribution of all the possible means of samples of a given size.

Sampling Error : Error or variation among sample statistics due to chance, that is, differences between each sample and the population, and among several samples, which are due solely to the elements we happen to choose for the sample.

Sampling Fraction : The fraction or proportion of the population contained in a sample.

Sampling with Replacement : A sampling procedure in which sampled items are returned to the population after being picked, so that some members of the population can appear in the sample more than once.

Sampling without Replacement : A sampling procedure in which sampled items are not returned to the population after being picked, so that no member of the population can appear in the sample more than once.

Scatter Diagram : A graph of points on a rectangular grid; the X and Y co-ordinates of each point correspond to the two measurements made on some particular sample element, and the pattern of points illustrates the relationship between the two variables.

Seasonal Variation : Patterns of change in a time series within a year; patterns that tend to be repeated from year to year.

Secular Trend : A type of variation in a time series, the value of the variable tending to increase or decrease over a long period of time.

Sensitivity Analysis : A technique for determining how the optimal solution to the linear program changes if the problem data change.

Significance Level : The probability of rejecting the null hypothesis when it is true.

Simple Random Sampling : Methods of selecting samples that allow each possible sample an equal probability of being picked and each item in the entire population an equal chance of being included in the sample.

Simplex Method : An efficient method for solving a linear programming problem.

Simultaneous Equations : When two or more equations are satisfied by the same values of the unknown quantities, they are referred to as simultaneous equations.

Skewness : The extent to which a distribution of data points is concentrated at one end or the other; the lack of symmetry.

Slack Variable : A variable used in linear programming to convert an inequality into an equation.

Slope : A constant for any given straight line whose value represents how much each unit change of the independent variable changes the dependent variable.

Special Cause Variation : Assignable or non-random variation is also referred to as special cause variation.

Standard Deviation : The positive square root of the variance.

Standard Error : The standard deviation of the sampling distribution of a statistic.

Standard Error of Estimate : A measure of the reliability of the estimating equation, indicating the variability of the observed points around the regression line, that is, the extent to which observed values differ from their predicted values on the regression line.

Standard Error of the Mean : The standard deviation of the sampling distribution of the mean.

Standard Error of the Regression Coefficient : A measure of the variability of sample regression coefficient around the true population regression coefficient.

Standard Normal Probability Distribution : A Normal Probability distribution, with mean $\mu = 0$ and standard deviation $\sigma = 1$.

State of Nature : A future event not under the control of the decision maker.

Statistical Dependence : The conditions when the probability of some event is dependent upon, or affected by, the occurrence of some other event.

Statistical Independence : The condition when the occurrence of one event has no effect upon the probability of occurrence of another event.

Statistics : Numerical measures describing the characteristics of a sample.

Strata : Groups within a population formed in such a way that each group is relatively homogeneous, but wider variability exists among the separate groups.

Stratified Sampling : A method of random sampling in which the population is divided into homogeneous groups, or strata, and elements within each stratum are selected at random according to one of two rules:

i. A specified number of elements is drawn from each stratum corresponding to the proportion of that stratum in the population or,

ii. Equal number of elements are drawn from each stratum, and the results are weighted according to the stratum's proportion of the total population.

Subjective Probability : Probabilities based on the personal beliefs of the person making the probability estimate.

Sufficient Estimator : An estimator that uses all the information available in the data concerning a parameter.

Symmetrical : A characteristic of a distribution in which each half is the mirror image of the other half.

Systematic Sampling : A method of random sampling used in statistics in which elements to be sampled are selected from the population at a uniform interval that is measured in time, order, or space.

Time Series : Information accumulated at regular intervals and the statistical methods used to determine patterns in such data.

Type I Error : Rejecting a null hypothesis when it is true.

Type II Error : Accepting a null hypothesis when it is false.

Unbiased Estimator : An estimator of a population parameter that, on the average, assumes values above the population parameter as often, and to the same extent, as it tends to assume values below the population parameter.

Unweighted Aggregates Index : Uses all the values considered, and assigns equal importance to each of these values.

Unweighted Average of Relatives Method : To construct an index number, this method finds the ratio of the current price to the base price for each product, adds the resulting percentage relatives, and then divides by the number of products.

Utility : The pleasure or displeasure someone derives from an outcome.

Variance : A measure of the average squared distance between the mean and each item in the population.

Venn Diagram : A pictorial representation of probability concepts in which the sample space is represented as a rectangle and the events in the sample space as portions of that rectangle.

Weighted Aggregates Index : Using all the values considered, this index assigns weights to these values.

Weighted Average of Relatives Method : To construct an index number, this method weighs by importance the value of each element in the composite.

Weighted Mean : An average calculated to take into account the importance of each value to the overall total, that is, an average in which each observation value is weighted by some index of its importance.

Y-Intercept : A constant for any given straight line whose value represents the value of the Y variable when the X variable has a value of 0.