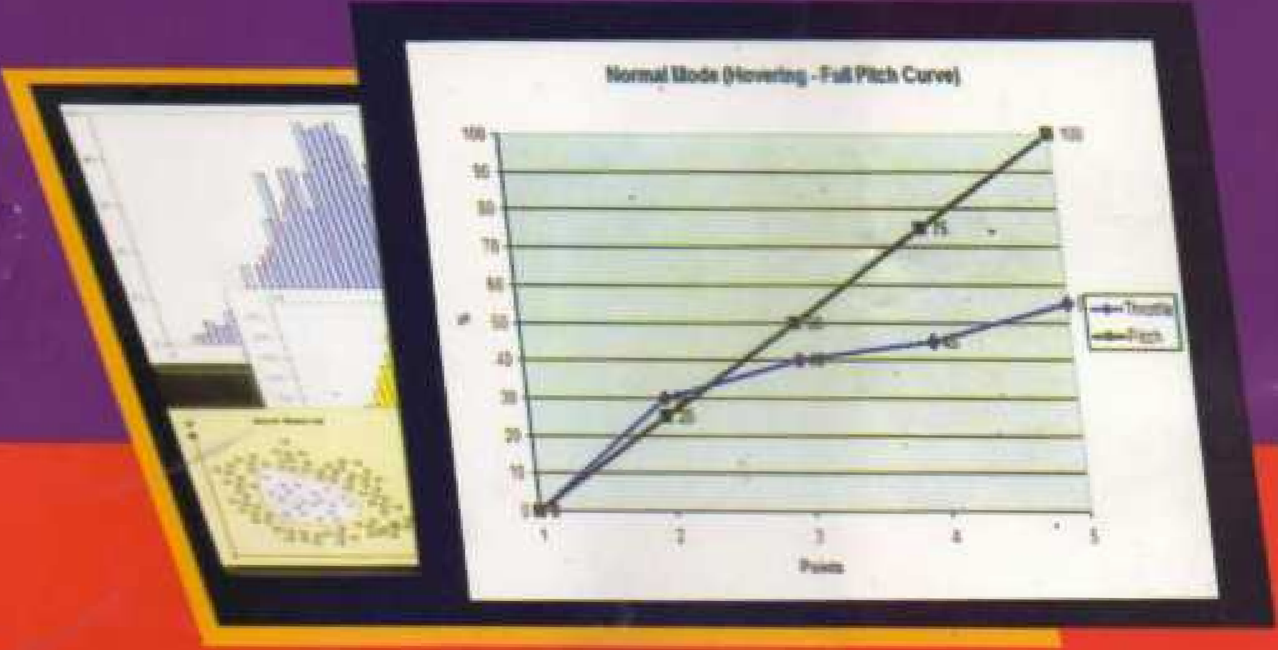
Part-I



ODISHA STATE BUREAU OF TEXTBOOK
PREPARATION AND PRODUCTION
PUSTAK BHAVAN, A-11, SUKA VIHAR, BHUBANESWAR

(Approved by the Council of Higher Secondary Education, Odisha, Bhubaneswar)

BOARD OF WRITERS

Dr. Jitendriya Sarangi
Retired Professor of Statistics
Utkal University

Dr. Udayanath Rout
Retired Head, Department of Statistics
Ravenshaw College (Autonomous)

Mrs. Pramoda Mohapatra
Retired Head, Department of Statistics
Kendrapara College (Autonomous)

Dr. Niranjan Mishra
Reader in Statistics
Rajendra College(Autonomous)

Dr. Kunja Behari Panda
Reader in Statistics
Utkal University

REVIEWER

Dr. Udayanath Rout
Retired Head, Department of Statistics
Ravenshaw College (Autonomous)

Published by

The Odisha State Bureau Of Textbook Preparation and Production,
Pustak Bhavan, Bhubaneswar

Published by

The Odisha State Bureau of Textbook Preparation and Production
Pustak Bhavan, Bhubaneswar, Odisha, India

First Edition : 2009/ 1000 Copies
Revised Edition : 2010/ 2000 Copies
Revised and Enlarged Edition : 2015
Publication No : 72
ISBN : 81- 8005 - 206 - 0

Type Setting : Sonali Prakashan

Printed at : Jagannath Process Pvt. Ltd.

Price ₹ : 197/- (Rupees One hundred Ninety Seven only)

The Odisha State Bureau of Textbook Preparation and Production has made a pioneering effort in publishing this textbook in Statistics which is expected to meet the requirement of students in the subject, both in the Science stream and Arts stream. The Bureau has utilised the best talents available within the state in preparation of this Text.

The present book is meant for the first year students of the +2 course. It has been authored by Prof. Jitendriya Sarangi, Prof. Udayanath Rout, Mrs. Pramoda Mohapatra, Dr. Niranjan Mishra and Dr. Kunja Behari Panda and reviewed by Prof. Udayanath Rout. The book has been written according to the latest revised syllabus of the Council of Higher Secondary Education, Odisha and gives a complete coverage of the subject matter.

I have every hope that the teachers and the students of Statistics will find this book useful and fruitful. Constructive suggestions for the improvement of this book shall be highly appreciated.

Director

Odisha State Bureau of Textbook Preparation and Production,
Pustak Bhavan, Bhubaneswar

tool for research in applied sciences. As such, for the academic interest of the students the subject is being taught at the +2 level by the Council of Higher Secondary Education (CHSE), Odisha. The Prescribed Syllabus has undergone several revisions and has been updated to meet the needs of the students intending to persue higher studies.

The present textbook viz. Statistics Part -1 has been written by a group of experienced teachers of the subject selected by the Bureau of Textbook Preparation and Production, Odisha according to the latest revised syllabus. The material of the book has been presented in simple language with lucid style along with a good number of illustrations and examples. Solved examples have been included in each chapter for easy understanding. A number of questions covering all topics have been included for solution. The efforts of the authors will be fruitful if the book is appreciated by the students and teachers as a fundamental book in the subject.

The authors are grateful to their colleagues, scholars of eminence and authors of similar books which have been very useful as guidelines in writing the text. The authors are also thankful to Prof. Jitendriya Sarangi and Dr. Udayanath Rout for the pains they have taken in reviewing the book.

The preparation of this text book was conceived by the State Text Book Preparation and Production, Department of Higher Education Odisha. Their constant assistance and support has matured in the publication of the book. The authors, indeed, are thankful to all the members of the staff of the Bureau for their keen interest and co-operation.

The authors might have overlooked some deficiencies in this text bok. Constructive criticisms and innovative feedbacks from the student readers, their esteemed teachers and others who have taken interest in this book are most welcome to improve this book in future.

Board of Writers

BINOMIAL THEOREM

EXPONENTIAL AND LOGARITHMIC SERIES

***

PERMUTATIONS AND COMBINATIONS

In this chapter we develop some techniques for counting the number of possible out comes of a particular experiment or the number of elements in a particular set. Such techniques are sometimes referred to as combinatorial analysis.

1.1. INTRODUCTION

Permutations refer to different arrangements of things from a given lot taken one or more at a time whereas combinations refer to different sets or groups made out of a given lot, without repeating an element, taking one or more of them at a time.

The distinction will be clear from the following illustration of combinations and permutations made out of a set of three elements {a, b, c}.

	Combinations	Permutations
(i)	One at a time: {a}, {b}, {c}	{a}, {b}, {c}
(ii)	Two at a time: {a,b}, {b,c}, {a,c}	{a,b}, {b,c}, {a,c}
		{b,a}, {c,b}, {c,a}
(iii)	Three at a time: {a,b,c}	{a,b,c} {a, c, b}
		{b,c,a} {b, a, c}
		{c,a,b} {c, b,a}

It may be noticed that on the left above, every set has different combination whereas on the right above, there are sets with different arrangements, wherever possible, of the same group. However, no element appears twice or more in any set, e.g., {a, a}, {b, b}, {c, c}, {a, b, b}, {c, c, c} etc.

1.2. FUNDAMENTAL PRINCIPLES OF COUNTING:

There are two fundamental rules of counting or selection based on the simple principles of addition and multiplication, the latter, when events occur independent of one only one operation can be done at a time.

This is known as addition principle of counting. Thus, since a king can be drawn from a pack of cards in 4 different ways (it may be a king of spade, club, heart or diamond) and a queen can also be drawn in 4 different ways, then either a king or a queen can be drawn in (4 +4) = 8 ways. This can be extended to several operations provided, only one operation can be done at a time.

(ii) Multiplication Principle:

If one operation can be done in 'm' different ways and when it has been done in anyone of the 'm' ways, a second operation can be done in 'n' different ways, then both the first and the second operations together can be done in (mxn) different ways. Further, if a third operation can be done in 'p' different ways, after completing the first and the second operations, the first, second and third operations together can be done in (m x n x p) different ways.

Similar results can be obtained for any number of operations done together or one after another.

Proof: Let $a_1, a_2, \ldots\ldots a_m$ be the 'm' ways of doing the first thing and $b_1, b_2, \ldots\ldots b_n$ be the 'n' ways of doing the second thing independent of the first. Then, the two things can be done simultaneously in the following ways:

$$\begin{array}{l} a_1b_1;\ a_1b_2;\ a_1b_3;\ldots\ldots\ldots\ldots;\ a_1b_n \\ a_2b_1;\ a_2b_2;\ a_2b_3;\ldots\ldots\ldots\ldots;\ a_2b_n \\ \cdot \quad \cdot \quad \cdot \qquad\qquad \cdot \\ \cdot \quad \cdot \quad \cdot \qquad\qquad \cdot \\ a_mb_1;\ a_mb_2;\ a_mb_3;\ldots\ldots\ldots\ldots;\ a_mb_n \end{array}$$

Thus, these are (m x n) number of ways of doing both the things simultaneously.

	1	2	3	4	5	6
1	**1, 1**	1, 2	1, 3	1, 4	1, 5	1, 6
2	2, 1	**2, 2**	2, 3	2, 4	2, 5	2, 6
3	3, 1	3, 2	**3, 3**	3, 4	3, 5	3, 6
4	4, 1	4, 2	4, 3	**4, 4**	4, 5	4, 6
5	5, 1	5, 2	5, 3	5, 4	**5, 5**	5, 6
6	6, 1	6, 2	6, 3	6, 4	6, 5	**6, 6**

Therefore, from the fundamental principle of counting, if repetitions are allowed, from N elements N elements can be taken in N^N ways. If , however, only 'r' of the N elements are taken at a time, the possible ways would be N^r, (6^2 in the above example). If repetitions are not allowed, the leading diagonal comprising {1,1}, {2,2} etc. is avoided and the total choices is 6 x 5 = 30 or n x (n - 1), only for two dice each with n faces.

Example:

Let a student has three different pants, say, P_1, P_2, P_3; four different shirts say S_1, S_2, S_3, S_4 and five different ties, say, T_1, T_2, T_3, T_4, T_5. Then he can go to the college with one pant, one shirt and one tie at a time in 3 x 4 x 5 = 60 different combinations as follows:

$P_1S_1T_1$, $P_2S_1T_1$, $P_3S_1T_1$, $P_1S_2T_1$, $P_2S_2T_1$, $P_3S_2T_1$, $P_1S_3T_1$, $P_2S_3T_1$, $P_3S_3T_1$, $P_1S_4T_1$, $P_2S_4T_1$, $P_3S_4T_1$, etc.

Note: Here addition principle can not be applied as it would mean, the student will be allowed to wear either a pant or a shirt or a tie, which he can do in 3+4+5 = 12 ways. This will lead to absurdity. One can not imagine a student going to the college by wearing a tie

(iii) the same route is not taken.

Solution:

(i) The man can go from A to B in 5 different ways, for he may take any one of the five routes. When he has done so in any of the 5 ways, he may return in any of the 5 different ways, i.e., there are 5 different ways of returning. So, the total number of different ways is 5 x 5 = 25.

(ii) While going, he chooses any one of the five routes in 5 ways. But while returning, he takes the same route in which he came. This can be done in 1 way. So, the total number of ways is 5 x 1 = 5 .

(iii) He can go from A to B in 5 different ways and return in 4 different ways. So the total number of ways is 5 x 4 = 20.

Example: 1.2

How many telephone connections can be allotted with 5 digits and 6 digits using the natural numbers 1 to 9 ?

Solution:

As per the rules of counting, the total number of telephone connections with five digits would be 9^5 = 59, 049 and with six digits would be 9^6 = 5, 31, 441

Example: 1.3

In how many ways can two persons A and B occupy their seats in a row of six seats ?

Solution:

Whoever comes first, he would be seated in 6 ways and after he occupies a seat, the other can be seated in 5 ways because one person can occupy only one seat. Therefore, both the persons A and B can be seated in 6 x 5 = 30 ways.

second person can choose his seat in 7 ways. The first and second persons can choose their seats in 8 x 7 different ways. The third person can choose a seat from the remaining six seats in 6 ways. Thus, the first, second and the third persons can choose their seats in 8 x 7 x 6 ways. Hence, the number of ways in which the three persons can choose their seats is 8 x 7 x 6 = 336.

1.3. FACTORIAL NOTATION:

The product of the positive integers from 1 to n oc :urs very often in mathematics and hence is denoted by the special symbol n! (read as "n factorial")

$$n! = 1.\ 2.\ 3.\ \ldots\ldots\ldots\ldots(n-2)(n-1)n$$

$$= n(n-1)(n-2)\ldots\ldots 3.2.1$$

$$= n[(n-1)!]$$

$$= n\ (n-1)\ [(n-2)!]$$

$$= n(n-1)(n-2)\ldots\ldots(n-r+1)\{(n-r)!\}$$

Hence $\frac{n!}{(n-1)!} = n$

Thus, $2! = 1.2 = 2,\quad 3! = 1.2.3 = 6,\quad 4! = 1.2.3.4 = 24$

$5! = 5.\ 4! = 5.\ 24 = 120,\quad 6! = 6.\ 5! = 6.\ 120 = 720$ etc.

Example: 1.5 Show that $\frac{10!}{8!} = 90$

Solution : $\frac{10!}{8!} = \frac{10.9.8!}{8!} = 90$

Example : 1.6 Show that $\frac{(n+1)!}{(n-2)!} = n^3 - n$

1.4. PERMUTATIONS:

(i) Each of the ordered arrangement which can be made by taking some or all of a number of distinct objects is called a permutation. For example, the permutations of the letters A, B, C, D taken two at a time are: AB, BA, AC, CA, AD, DA, BC, CB, BD, DB, CD and DC.

Thus, there are 12 permutations or different arrangements when two objects are taken at a time out of 4 objects.

We can obtain six numbers by arranging three digits, 5,6 and 7. The six arrangements or permutations are,

567, 576, 657, 675, 756 and 765.

(ii) An arrangement of a set of 'n' objects in a given order is called a permutation of the objects (taken all at a time). An arrangement of any $r \leq n$ of these objects in a given order is called an r-permutation of n objects or a permutation of 'n' objects taken 'r' at a time.

The number of permutations of 'n' objects taken 'r' at a time, where $r \leq n$ is denoted by ^nP_r or P(n, r).

Example: 1.7

Find the number of permutations of 6 objects, say a, b, c, d, e, f taken three at a time.

Solution:

Let the selected three letters form a word and be represented by three boxes.

□ □ □

Now the first letter can be chosen in 6 different ways; following this, the second letter can be chosen in 5 different ways; and , following this, the last letter can be chosen in 4 different ways.

given in the preceding example. The first element in an r-permutation of n-objects can be chosen in 'n' different ways; following this, the second element in the permutation can be chosen in n-1 ways; and, following this, the third element in the permutation can be chosen in n- 2 ways. Continuing in this manner , we have that the rth (last) element in the r-permutation can be chosen in $n-(r-1) = n-r+1$ ways.

Thus, $P(n, r) = {}^{n}P_{r} = n(n-1)\,(n-2)\ \ldots$.to r factors

$$= n(n-1)\,(n-2)\ \ldots(n-r+1) \tag{1.1}$$

Note:

1. The number of permutations of 'n' different things taken all at a time is

$${}^{n}P_{n} = n(n-1)(n-2)\ \ldots 3.2.1 = n!$$

2. $${}^{n}P_{n-1} = n(n-1)(n-2)\ \ldots 3.2.1 = {}^{n}P_{n}$$

3. $${}^{n}P_{r} = n(n-1)(n-2)\ \ldots(n-r+1)$$

$$= \frac{n(n-1)(n-2)\ldots(n-r+1)\{(n-r)!\}}{(n-r)!} = \frac{n!}{(n-r)!}$$

4. We have ${}^{n}P_{n} = n!$

Also $${}^{n}P_{n} = \frac{n!}{(n-n)!} = \frac{n!}{0!}$$

$$\Rightarrow \quad n! = \frac{n!}{0!}$$

Hence, $$0! = \frac{n!}{n!} = 1$$

LOGARITHMS. (The words may not have any meaning).

Solution:

There are 10 different letters. Therefore, 'n' is equal to 10 and since we have to find four-letter words, 'r' is 4. Hence, the required number of words is

$$^{10}P_4 = \frac{10!}{(10-4)!} = \frac{10!}{6!} = \frac{10 \times 9 \times 8 \times 7 \times 6!}{6!} = 10 \times 9 \times 8 \times 7 = 5{,}040.$$

Example: 1.9

How many four digit numbers, each greater than 7,000, can be formed from the digits 3,5,7,8 and 9 using each digit once?

Solution:

As the numbers are to be greater than 7000, the digit of the thousand place can be any one of the digits 7,8 and 9. Now the thousand place digit can be chosen in 3 ways, (since $^3P_1 = 3$) and the remaining three digits of the hundred, ten and unity places can be any of the four digits left, which can be chosen in $^4P_3$ ways. Therefore, the total number of ways is $3 \times {}^4P_3 = 3 \times 4 \times 3 \times 2 = 72$.

Example: 1.10

In how many ways can 5 English, 3 Odiya and 3 Hindi books be arranged so that the books of the same language are kept together?

Solution:

Each language book amongst themselves can be arranged in the following ways.

English : 5 books in $^5P_5$ i.e., 5! ways.

Odiya : 3 books in $^3P_3$ i.e., 3! ways.

(b) Circular Permutations:

(i) Circular permutations are similar to the arrangements of objects along the circumference of a circle. Here, there is neither a beginning nor an end. We fix the position of one object and then arrange the remaining (n – 1) objects in all possible ways. This can be done in (n – 1)! ways.

(ii) In the above arrangements, clockwise and anticlockwise order of arrangements have been distinguished. If this distinction is not made, the number of arrangements would be $\frac{1}{2}$(n – 1)!. We distinguish between these two in the following two examples.

Example: 1.11

In how many ways can 5 boys and 5 girls be seated around a table so that no two boys are adjacent ?

Solution :

Let the girls be seated first. They can sit in 4! ways. Now, since the places for the boys between girls are fixed, the boys will occupy the remaining 5 places. There are 5! ways for the boys to fill up the 5 places in between 5 girls already seated around a table. Thus, the total number of ways in which both the girls and the boys can be seated is

4! x 5! = 2880 ways.

Example: 1.12

If 7 persons are seated around a table so that in all of the arrangements the same neighbours are not present, then the required number of ways will be $\frac{1}{2}$(n -1)! i.e. $\frac{1}{2}$. 6!

$= \frac{1}{2}$. 6.5.4.3.2.1 = 360

$p!.q!.r!$

Proof: Let the 'n' things be such that p of them are of one kind denoted by a, q of them are of second kind denoted by b, r of them are of third kind denoted by c and all the rest are different.

Let 'x' be the required number of permutations. If the p a's be replaced by p new letters, $a_1, a_2, \ldots a_p$ which are different from each other and different from the rest, then without changing the position of any other letter , they would produce p! permutations. If this change be made in each of the x permutations, this would produce x . p! permutations i.e, the total number of permutations would become x . p!. Again, if the 'q' b's be replaced by q new letters, $b_1, b_2, \ldots b_q$ different from each other and different from the rest, then the total number of permutations would become x . p! . q!.

Again, if the 'r' c's be replaced by r new letters $c_1, c_2, \ldots c_r$ different from each other and different from the rest, the total number of permutations would become x . p! . q! . r!.

But now the n things are all different and the permutations of n different things taken all at a time is n!.

$$\therefore \quad x.p!.q!.r! = n! \;\Rightarrow\; x = \frac{n!}{p!.q!.r!}$$

The above principle can easily be generalized..

Example:

(a) The letters of the word 'ALLAHABAD' can be arranged in

$$\frac{9!}{4!\ 2!} = \frac{9 \times 8 \times 7 \times 6 \times 5}{2} = 7560 \text{ ways.}$$

Here n = 9, p = 4 (since there are four A's), q = 2 (as L appears twice) and the rest of the letters H, B and D appear once each.

Solution:

The word ACCOUNTANT has 10 letters of which 2 are 'A's, 2 are 'C's , 2 are ,'N's and 2 are 'T's, the rest are different . Therefore the number of permutations is

$$\frac{10!}{2!2!2!2!} = \frac{10.9.8.7.6.5.4.3.2.1}{2.2.2.2} = 2,26,800$$

Example: 1.14

How many numbers greater than a million can be formed with the digits 4,5,5,0,4,5,3?

Solution:

Each number must consist of 7 or more digits. There are 7 digits in all, of which there are 2 fours, 3 fives and the rest different.

So, the total number is $\frac{7!}{3!\ 2!} = 420$.

Of these numbers, some begin with zero and are less than one million which must be rejected.

The total number of such numbers beginning with zero is $\frac{6!}{2!\ 3!} = 60$.

Hence the required number is 420 – 60 = 360.

(d) Permutations with Repetitions:

The number of permutations of 'n' different objects taken 'r' at a time when each object may be repeated any number of times in any permutation is 'n^r'.

Suppose we have four digits 2,4,6,8 and we have to form four digit numbers when each digit may be repeated any number of times. This can be done in 4^4 = 256 ways. Thus there can be 256 such numbers . These numbers will include those numbers where, 8882, 4666,6668, 2444 etc.

(iv) one digit is repeated four times, such as

2222, 4444,6666, 8888

Further, four books can be given to five students in $5^4 = 625$ ways, when no restriction is placed on the distribution of the books. The first book can be given to any of the 5 students in 5 ways. The second book can also be given to any of the 5 students in 5 ways. Similarly, the third and fourth books can be disposed of in 5 ways each. Hence all the four books can be distributed in $5 \times 5 \times 5 \times 5 = 5^4 = 625$ ways. Here all possible types of distributions are included. Similarly, 6 ball pens can be placed in three packets in $3^6 =$ 729 ways.

(e) Restricted permutations:

(i) The number of permutations of 'n' different things taken 'r' at a time in which 'p' particular things do not occur is $^{n-p}P_r$

Keep aside the p particular things not to occur and then arrange the remaining n - p things in r places. This can be done in $^{n-p}P_r$ ways.

(ii) The number of permutations of 'n' different things taken 'r' at a time in which 'p' particular things are always present is $^{n-p}P_{r-p} \times {}^{r}P_p$

Keep aside the 'p' particular things and form the permutations of the remaining n - p things taken r - p at a time. The number of such permutations is $^{n-p}P_{r-p}$

In each of these permutations introduce the 'p' particular things kept aside, one by one. The first thing can be introduced in r – p +1 ways. After introducing the first, the second thing can be introduced in r – p +2 ways and the pth thing in r – p +p or 'r' ways.

$$^{8-3}P_4 = {}^5P_4 = \frac{5!}{(5-4)!} = 120 \text{ ways.}$$

In case, digits 2,4 and 6 are always included, then the required number of 4 -digit numbers will be

$$^{8-3}P_1 . {}^4P_3 = {}^5P_1 \times {}^4P_3 = 120$$

Note: Here we have assumed that no digit is allowed to repeat itself.

Example: 1.15

In how many ways can the letters of the word 'STRANGE' be arranged so that

(i) the vowels are never separated.

(ii) the vowels never come together, and

(iii) the vowels occupy only the odd places.

Solution:

(i) There are 7 letters. Since the vowels are not to be separated we may regard them as forming one letter. So there are six letters S,T, R, N, G and (A E). They can be arranged among themselves in 6! ways. The two vowels can again be arranged in 2! ways. So, the total number of arrangements = 6! x 2! = 1440.

(ii) The number of arrangements in which the vowels do not come together can be obtained by subtracting from the total number of arrangements the number of arrangements in which the vowels come together. Since the total number of arrangements is 7! and the number of arrangements in which the vowels come together is 6! x 2!, the number of arrangements in which the vowels do not come together is

$$7! - 6! \times 2! = 6! \times 5 = 3600$$

arranging the vowels can be associated with each of the ${}^{5}P_{5}$ ways of arranging the consonants.

So, the total number of arrangements = ${}^{4}P_{2}$ x ${}^{5}P_{5}$ = 12 x 120 = 1440

Example: 1.16

How many integers of six digits can be formed from the digits 4,5,6,7,8,9; no digit being repeated? How many of them are not divisible by 5?

Solution:

The six digits all being different, can be arranged among themselves in 6! ways. Let us find the number of integers divisible by 5. The integers in which 5 occurs in the unit place are divisible by 5. Fix 5 at the unit place. The remaining five digits can be arranged among themselves in 5! ways. So, the number of integers divisible by 5 is 5!. Hence, the number of integers which are not divisible by 5 is 6! - 5! = 600.

1.5. COMBINATIONS:

(a) Each of the group or selection which can be made by taking some or all of a number of objects without reference to order of the objects in each group is called a combination.

For example , the combinations of the letters A, B, C, D taken two at a time are AB, AC, AD, BC, BD and CD. Thus, there are 6 combinations or groups or selections which can be made when two objects are taken at a time out of four objects.

In combination, order does not matter. Thus AB and BA mean the same combination. A change in any one object will constitute a new combination. Thus AB and AC are two different combinations.

To denote the number of combinations of 'n' different objects taken 'r' at a time we use the symbol ${}^{n}C_{r}$ or C(n, r) or $\binom{n}{r}$.

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

Hence we can write $\quad ^{n}C_{r} = \frac{^{n}P_{r}}{r!}$

Theorem:

The number of combinations of 'n' different things taken 'r' at a time is given by

$$^{n}C_{r} = \frac{n!}{r!(n-r)!}, \text{ where } (r \leq n) \qquad (1.2)$$

Proof:

Let $^{n}C_{r}$ denote the required number of combinations of 'n' different things taken 'r' at a time. Each of these combinations has 'r' different things.

So, if the 'r' different things be arranged among themselves in all possible ways, each combination would produce r! permutations. Hence, $^{n}C_{r}$ combinations would produce $^{n}C_{r}$ x r! permutations. But this number is clearly equal to the number of permutations of 'n' different things taken 'r' at a time. Hence $^{n}C_{r}$ x r! = $^{n}P_{r}$

$$\Rightarrow \; ^{n}C_{r} = \frac{^{n}P_{r}}{r!} = \frac{n(n-1)(n-2).....(n-r+1)}{r!}$$

$$= \frac{n(n-1)(n-2).....(n-r+1)}{r!} \cdot \frac{(n-r)!}{(n-r)!} = \frac{n!}{r!(n-r)!}$$

(iii) $\quad {}^nC_n = 1$

Proof: $\quad {}^nC_n = \dfrac{n!}{n!(n-n)!} = \dfrac{n!}{n!o!} = \dfrac{n!}{n!\times 1} = 1$

(iv) $\quad {}^nC_{n-1} = n$

(v) $\quad {}^nP_r = r!\times {}^nC_r$

(vi) $\quad {}^nC_r$ is also called the 'r' combination of 'n' different objects.

(vii) $\quad {}^nC_r = \dfrac{n}{r}\times {}^{n-1}C_{r-1}$

Proof: Let us find the number of 'r' combinations of 'n' different objects in which a particular object say a_1 would always occur. The number of such combinations is ${}^{n-1}C_{r-1}$. Similarly the number of 'r' combinations of 'n' different objects containing a_2 is ${}^{n-1}C_{r-1}$, the number of 'r' combinations of 'n' different objects containing a_3 is also ${}^{n-1}C_{r-1}$ and so on. The total number of such objects is $n_,$. Therefore the total number of letters written in these combinations would be (${}^{n-1}C_{r-1} + {}^{n-1}C_{r-1}$ +..... n times) = n x ${}^{n-1}C_{r-1}$

But the number of objects in 'r' combinations of 'n' different objects is r. ${}^nC_r$.

$$r\times {}^nC_r = n\times {}^{n-1}C_{r-1} \Rightarrow \quad {}^nC_r = \frac{n}{r}\times {}^{n-1}C_{r-1}$$

This result is true for all integral values of r and n.

Note - This can be proved by using factorial notations.

Complementary Theorem:

The number of combinations of 'n' different things taken 'r' at a time, is same as the number of combinations of 'n' different things taken (n - r) at a time, i.e.,

$${}^nC_r = {}^nC_{n-r}, \text{ where } 0 \le r \le n.$$

(i) If ${}^nC_r = {}^nC_p$ then either r = p or r+p = n. This is because ${}^nC_r = {}^nC_{n-r}$, and so

n - r = p or r + p = n

(ii) If in the formula ${}^nC_{n-r} = {}^nC_r$, we put r = n then ${}^nC_0 = {}^nC_n = 1$.

So we consider the value of ${}^nC_0$ equal to 1.

(iii) If in (ii) above, we put r = n - 1, then ${}^nC_1 = {}^nC_{n-1} =$ n etc.

If in the formula ${}^nC_r = \frac{n!}{r!(n-r)!}$ we put r = n, we find ${}^nC_n = \frac{n!}{n!o!}$. But as ${}^nC_n = 1$; $\frac{n!}{n!o!} = 1$ is valid only if 0! is equivalent to 1. Thus the value of 0! in combinatorial analysis is considered equivalent to 1, though strictly speaking, it has no meaning.

(iv) $\frac{{}^nC_r}{{}^nC_{r-1}} = \frac{n-r+1}{r}$ · This is because,

$$\frac{{}^nC_r}{{}^nC_{r-1}} = \frac{n!}{r!(n-r)!} \cdot \frac{(r-1)!(n-r+1)!}{n!}$$

$$= \frac{(r-1)!(n-r+1).(n-r)!}{r.(r-1)!(n-r)!} = \frac{n-r+1}{r}$$

(c) Total number of combinations:

(i) The total number of combinations of 'n' different things taken some or all at a time is $2^n - 1$.

Let the 'n' different things be denoted by $a_1, a_2, \ldots, a_n$. While making a selection of some taken is n factors = 2^n ways. But this includes the case in which none of the things has been taken. Rejecting this case, the total number of combinations would be $2^n - 1$.

Note: The total number of combinations of 'n' different things taken some or all at a time is $^nC_1 + {}^nC_2 + \dots\dots\dots + {}^nC_n$. So, we have

$$^nC_1 + {}^nC_2 + \dots\dots\dots + {}^nC_n = 2^n - 1$$

Thus, we can select some or all of the 6 players at a time in $2^6 - 1 = 64 - 1 = 63$ ways.

(ii) The total number of combinations of (p+q+r+....) things, where 'p' are alike of one kind, 'q' are alike of second kind, 'r' are alike of the third kind and so on, taken any number at a time is (p+1)(q+1)(r+1) ...... –1

Consider the 'p' things which are alike. The 'p' things can be dealt with in (p+1) ways, for we may take 1 or 2 or 3... or 'p' or none in any selection. Similarly the 'q' alike things can be dealt with in (q+1) ways, 'r' alike things in (r+1) ways etc. Since each way of assigning one can be associated with each way of the others the total number of dealing with them is (p+1)(q+1)(r+1).........

But this number includes one case where all things are left out. Therefore, the total number of ways is (p+1)(q+1)(r+1).........–1

Example: 1.17

In order to pass in C. A. Intermediate Examination, a student has to secure minimum pass marks in each of the 7 subjects. In how many ways can a student fail?

Solution:

Each subject can be dealt with in two ways; the student may pass or fail in it. So the 7 subjects can be dealt in 2^7 ways. But this includes the case in which the student passes in all the 7 subjects. Rejecting this case, the number of ways in which the student can fail is $2^7 - 1 = 127$.

contain all different letters, some may not contain all different letters. Following cases arise.

(a) All the four letters are different.

(b) Two letters are alike, two others are different.

(c) Two letters are alike of one kind, two are alike of other kind.

(a) There are 8 different letters. The required number of combinations = ${}^{8}C_{4}$

(b) There are three pairs of alike letters, i.e., (O,O), (I, I), (N, N). One pair can be chosen in ${}^{3}C_{1}$ ways. Remaining two different letters can be selected from remaining seven different letters in ${}^{7}C_{2}$ ways. Hence the number of combinations of this type is ${}^{3}C_{1} \times {}^{7}C_{2}$.

(c) Two pairs of similar letters can be chosen in ${}^{3}C_{2}$ ways.

Hence the total number of required combinations is

$${}^{8}C_{4} + ({}^{3}C_{1} \times {}^{7}C_{2}) + {}^{3}C_{2} = 136$$

Example: 1.19

Out of 10 consonants and 4 vowels, how many words can be formed each containing 6 consonants and 3 vowels?

Solution:

6 consonants can be chosen out of 10 in ${}^{10}C_{6}$ ways and 3 vowels can be chosen out of 4 in ${}^{4}C_{3}$ ways. So, the number of selections is ${}^{10}C_{6} \times {}^{4}C_{3}$

Each of these selection contains 9 letters which can be arranged among themselves in 9! ways.

So, the total number of words is ${}^{10}C_{6} \times {}^{4}C_{3} \times 9! = 304819200$.

the other side two. So, one must be placed on the side of three men and the remaining two on the other side. This can be done in ${}^3C_1$ or 3 ways. Again, 4 men on each side can be arranged among themselves in 4! ways. Hence the required number of ways is

$$ {}^3C_1 \times 4! \times 4! = 3 \times 24 \times 24 = 1728. $$

Example: 1.21 Find n, if ${}^nC_6 : {}^{n-3}C_3 = 91 : 4$

Solution: $${}^nC_6 = \frac{n!}{6!(n-6)!} \text{ and } {}^{n-3}C_3 = \frac{(n-3)!}{3!(n-3-3)!}$$

$$\therefore \quad \frac{{}^nC_6}{{}^{n-3}C_3} = \frac{n!}{6!(n-6)!} \times \frac{3!(n-6)!}{(n-3)!}$$

$$= \frac{n(n-1)(n-2)}{6.5.4}$$

So, $$\frac{n(n-1)(n-2)}{6.5.4} = \frac{91}{4}$$

$$\therefore \quad n(n-1)(n-2) = 5.\,6.91 = 5.6.7.13 = 15.14.13$$

Expressing the R.H.S as the product of three consecutive integers in descending order, we get n = 15.

Example: 1.22

A party of 6 is to be formed from 10 boys and 7 girls consisting of 3 boys and 3 girls. In how many different ways can the party be formed, if two particular girls refuse to join the same party together ?

Solution:

If the two particular girls do not refuse to join the same party, then we can select 3 girls from 7 in ${}^7C_3$ ways and 3 boys from 10 in ${}^{10}C_3$ ways. Hence a party of 6 including

$^{5}C_{1} \times {}^{10}C_{3} = 5 \times 120 = 600$ ways ....(2)

We notice that in the arrangements (2), those two particular girls who refuse to join are included. Hence the required number of arrangements can be obtained by subtracting (2) from (1) i.e., 4200 – 600 = 3600

(d) Some results:

(i) Combination of objects where some are alike and ʳome are different:

The total number of combinations of (p+q+r) objects of which 'p' are alike and of one kind, 'q' are alike and of another kind and the remaining 'r' are all different is given by

$(p + 1)(q + 1)2^{r} - 1$

(ii) Division into groups:

The number of ways in which (m+n+p) objects can be divided into 3 groups of m, n and p objects respectively is given by (m+n+p)! / m! n! p!

Thus 22 objects can be divided into 4 groups of 4,5,6 and 7 objects in $\frac{22!}{4!5!6!7!}$ ways.

(iii) Division into equal groups :

The number of ways in which '3n' objects can be divided into 3 equal but distinct groups of 'n' objects each is given by $(3n)! / (n!)^{3}$.

If no distinction is made between the 3 groups of 'n' objects each, then the number of ways in which (3n) objects can be divided into 3 equal groups of 'n' objects each is given by $(3n)! / 3!(n!)^{3}$.

(e) Restricted Combinations:

(i) The number of combinations of 'n' things taken 'r' at a time in which 'p' particular things always occur is $^{n-p}C_{r-p}$

never occur is $^{n-p}C_r$.

Let the 'p' particular things which never occur be set aside. Then there will remain (n - p) things out of which 'r' things may be selected in $^{n-p}C_r$ ways.

In none of these selections the 'p' particular things will occur. Hence the required number of combinations is $^{n-p}C_r$.

Example: 1.23 Prove that $^{n+1}C_r = {}^nC_r + {}^nC_{r-1}$

Solution: We know that $^nC_r = \dfrac{n!}{r!(n-r)!}$

RHS: $$^nC_r + {}^nC_{r-1} = \frac{n!}{r!(n-r)!} + \frac{n!}{(r-1)!(n-r+1)!}$$

$$= \frac{n!}{r(r-1)!(n-r)!} + \frac{n!}{(r-1)!(n-r+1)(n-r)!}$$

$$= \frac{n!}{(r-1)!(n-r)!}\left[\frac{1}{r} + \frac{1}{n-r+1}\right] = \frac{n!}{(r-1)!(n-r)!}\left[\frac{n+1}{r(n-r+1)}\right]$$

$$= \frac{(n+1)!}{r!(n-r+1)!} = {}^{n+1}C_r$$

Example: 1.24 Find the value of r if $^{18}C_r = {}^{18}C_{r+2}$

Solution: We know that $^nC_r = {}^nC_{n-r}$

So, $^{18}C_r = {}^{18}C_{18-r}$

Hence, $^{18}C_{18-r} = {}^{18}C_{r+2}$

$\Rightarrow 18 - r = r + 2$

$\Rightarrow 2r = 16$

$\Rightarrow r = 8.$

questions?

Solution:

(i) 8 questions out of 10 can be selected in $\begin{pmatrix}10\\8\end{pmatrix} = \begin{pmatrix}10\\2\end{pmatrix} = \frac{10.9}{1.2} = 45$ ways.

(ii) If he answers the first 3 questions , then he can choose the other 5 questions from the remaining 7 questions in $\begin{pmatrix}7\\5\end{pmatrix} = \begin{pmatrix}7\\2\end{pmatrix} = \frac{7.6}{1.2} = 21$ ways.

(iii) If he answers all the first 5 questions, then he can choose the other 3 questions from the last 5 in $\begin{pmatrix}5\\3\end{pmatrix} = 10$ ways. On the other hand, if he answers only 4 of the first 5 questions, then he can choose these 4 in $\begin{pmatrix}5\\4\end{pmatrix} = \begin{pmatrix}5\\1\end{pmatrix} = 5$ ways, and he can choose the other 4 questions from the last 5 in $\begin{pmatrix}5\\4\end{pmatrix} = \begin{pmatrix}5\\1\end{pmatrix} = 5$ ways;

Hence, he can choose the 8 questions in 5.5 = 25 ways. Thus, he has a total of 10 + 25 = 35 choices.

Example: 1.26

In how many ways can 7 toys be divided among 3 children if the youngest gets 3 toys and each of the other gets 2?

Solution:

Here we have to distribute 7 different objects in three distinct categories containing 3,2,2 objects. This can be done in $\frac{7!}{3!2!2!} = 210$ ways.

1. An automobile dealer provides motor cycles and scooters in two body patterns and five different colours each. Indicate the number of choices open to a customer visiting him.

2. Show that $30! = 2^{15} . 15! . (1.3.5........29)$

3. Simplify: (i) $\dfrac{n!}{(n-1)!}$ (ii) $\dfrac{(n+2)!}{n!}$

4. Compute: (i) $\dfrac{13!}{11!}$ (ii) $\dfrac{7!}{10!}$

5. In how many ways can 4 Indians and 4 Pakistanis be seated at a round table so that no two Indians sit together?

6. The Chief Ministers of 18 States in India meet to discuss the problems of unemployment. In how many ways can they seat themselves at a round table if the Odisha and Andhra Pradesh chief ministers choose to sit together?

7. If repetitions are not permitted.

(i) how many 3 digit numbers can be formed from the six digits 2,3,5,6,7 and 9?

(ii) how many of these are less than 400?

(iii) how many are even?

(iv) how many are odd?

(v) how many are multiples of 5?

8. In how many ways can a party of 7 persons arrange themselves

(i) in a row of 7 chairs?

(ii) around a circular table?

so that those of the same nationality sit together?

(ii) Solve the problem (i) if they sit at a round table?

11. Find 'n' if

(i) $P(n,2) = 72$, (ii) $P(n,4) = 42P(n,2)$ (iii) $2P(n,2) + 50 = P(2n, 2)$

12. Find the number of permutations of letters in the word 'ENGINEERING'.

13.(i) How many different words can be made with all the letters in the word 'ALLAHABAD'?

(ii) In how many of those will the vowels occupy the even places ?

14. If ${}^nP_4 = 12. {}^nP_2$, find n.

15. Find the value of 'n' if four times the number of permutations of 'n' things taken 3 together is equal to five times the number of permutations of (n - 1) things taken 3 together.

16. Prove that ${}^nP_r = n \times {}^{n-1}P_{r-1}$

17. How many numbers less than 1000 and divisible by 5 can be formed with digits 0,1,2,3,4,5,6,7,8,9 such that a digit does not occur more than once in a number ?

18. In how many different ways can 8 examination papers be arranged in a line so that the best and the worst papers are never together?

19. In how many ways can 3 boys and 5 girls be arranged in a row so that all the 3 boys are together ?

20. A number of four different digits is formed by using the digits 1,2,3,4,5,6,7 in all possible ways. Find

(i) how many such numbers can be formed ? and

(ii) how many of them are greater than 3400?

repetitions of the digits 1,2,5,6,8?

(ii) How many even numbers each lying between 100 and 1000 can be formed without repetitions of the digits 3,4,5,6,7?

24. Fill in the blanks:

(i) If $\frac{(2n+1)!}{(n+2)!}\cdot\frac{(n-1)!}{(2n-1)!}=\frac{3}{5}$, then n= ....

(ii) There are 10 true -false questions in an examination. The questions can be answered in .... ways.

(iii) The number of ways in which the letters of the word "TRIANGLE" can be arranged such that the three vowels are together, is ....

(iv) By using the digits 1,2,5,5,4 we can form............. even five digit numbers.

(v) The number of diagonals that can be drawn by joining the vertices of a hexagon is ....

(vi) Rakesh has five friends. He can invite one or more of them to a dinner party in ....ways.

25. Some statements are given below. If the statement is correct, write True otherwise write False.

(i) $\frac{(2k)!}{k!}=2^k\left[1.3.5.....(2k-1)\right]$

(ii) There are six multiple choice questions in an examination. If the first three questions have 4 choices each and the next three have 5 choices each, the questions can be answered in 8000 ways (all the answers may not be correct).

(iii) 1440 words can be formed each of 3 vowels and 2 consonants from the letters of the word 'INVOLUTE'.

(i) If $^{56}P_{r+6} : {}^{54}P_{r+3} = 30800 : 1$, then r = ...

(a) 31 (b) 41 (c) 51 (d) None of these

(ii) From 6 boys and 7 girls a committee of 5 is to be formed so as to include at least one girl. The number of ways in which this can be done is

(a) $^{13}C_4$ (b) $^{6}C_4 + {}^{7}C_1$ (c) $7 \times {}^{6}C_4$ (d) None of these

(iii) The number of ways in which 18 different books can be divided equally among 3 students, is

(a) $\dfrac{18!}{(6!)^3}$ (b) $\dfrac{18!}{3(6!)^3}$ (c) $\dfrac{3.18!}{(6!)^3}$ (d) None of these

(iv) There are three prizes to be distributed among five students. If no student gets more than one prize, then this can be done in

(a) 10 ways (b) 30 ways (c) 60 ways (d) None of these

(v) A professor has 8 scholars working under him. He takes them 3 at a time to conferences as often as he can without taking the same 3 scholars more than once. How many times can the professor go ?

(a) 21 times (b) 56 times (c) 45 times (d) None of these

27. For an examination, a candidate has to select 7 subjects from three different groups A, B and C. The three groups contain 4,5,6 subjects respectively . In how many different ways can a candidate make his selection if he has to select at least 2 subjects from each group ?

31. In how many ways can 12 students be partitioned into 3 teams A_1, A_2 and A_3, so that each team contains 4 students?

32. Find the number of ways in which

(i) a selection. (ii) an arrangement

of 4 letters can be made from the letters of the word 'MATHEMATICS'

33. Out of 5 males and 6 females, a committee of 5 is to be formed. Find the number of ways in which it can be done so that among the persons chosen in the committee, there are

(i) 3 males and 2 females

(ii) 2 males

(iii) no female

(iv) at least one female

(v) not more than 3 males.

34. Ten couples attended a party from whom six persons were chosen for a game. In how many ways can this be done so as to include exactly one couple?

35. Prove that $\frac{{}^nP_1}{1!}+\frac{{}^nP_2}{2!}+\frac{{}^nP_3}{3!}+......+\frac{{}^nP_n}{n!}=2^n-1$

***

two terms, viz. $(a + b)$, $(x - a)$, $(4x^2 + 3y^2)$, etc.

From elementary algebra, we know

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

$(a + b)^3 = (a + b)(a + b)^2 = (a + b)(a^2 + 2ab + b^2) = a^3 + 3a^2b + 3ab^2 + b^3$

These are quite simple to compute, but if the expansion contains a power (called index) which is of higher order or is negative or fractional, the expansion of such an expression by way of multiplication becomes tedious and complicated. Such expansions can be made by using an algebraic formula called Binomial Theorem.

1.6.1. Binomial Theorem for a Positive Integral Index :

Statement :

If $(x + a)$ is a binomial expression, the expansion of $(x + a)^n$ is given by

$$(x+a)^n = x^n + {}^nC_1 x^{n-1} a + {}^nC_2\, x^{n-2} a^2 + {}^nC_3\, x^{n-3} a^3 + \ldots\ldots + {}^nC_r\, x^{n-r} a^r + \ldots\ldots + {}^nC_n\, a^n$$

where 'n' is a positive integer.

Proof : We give below the proof of the theorem by the method of induction.

Step-I : By actual multiplication, we have,

$(x+a)^2 = x^2 + 2ax + a^2 = x^2 + {}^2C_1 x a + {}^2C_2\, a^2$

$(x+a)^3 = x^3 + 3x^2a + 3xa^2 + a^3 = x^3 + {}^3C_1\, x^2 a + {}^3C_2\, xa^2 + {}^3C_3\, a^3$

Thus the theorem is true when n has the values 2 and 3.

Step-II : To prove this theorem by the principle of mathematical induction we shall assume that the theorem is true for some particular value, say, m of n and we shall show that it is true for the value m + 1 of n also.

$$+ {}^mC_2\, x^{m-2}a^2 + \ldots + {}^mC_r\, x^{m-r}a^r + \ldots\ldots + {}^mC_m\, a^m]$$

$$= (x^{m+1} + {}^mC_1 x^m\, a + {}^mC_2\, x^{m-1}a^2 + \ldots + {}^mC_r\, x^{m-r+1}a^r + \ldots\ldots + {}^mC_m\, x\, a^m) + (x^m a$$

$$+ {}^mC_1 x^{m-1} a^2 + {}^mC_2\, x^{m-2}a^3 + \ldots + {}^mC_r\, x^{m-r}a^{r+1} + \ldots\ldots + {}^mC_m\, a^{m+1})$$

$$\therefore \quad (x+a)^{m+1} = x^{m+1} + (1 + {}^mC_1\,)x^m\, a + ({}^mC_1 + {}^mC_2)\; x^{m-1}a^2 + \ldots + ({}^mC_{r-1} + {}^mC_r)$$

$$x^{m-r+1}a^r + \ldots\ldots + {}^mC_m\, a^{m+1}$$

Using the relation ${}^mC_{r-1} + {}^mC_r = {}^{m+1}C_r$ and ${}^mC_m = 1 = {}^{m+1}C_{m+1}$ we have,

${}^mC_1 + 1 = m + 1 = {}^{m+1}C_1,\ {}^mC_1 + {}^mC_2 = {}^{m+1}C_2$, etc

So, $(x+a)^{m+1} = x^{m+1} + {}^{m+1}C_1\, x^m\, a + {}^{m+1}C_2\, x^{m-1}a^2 + \ldots + {}^{m+1}C_{m+1}\, a^{m+1}$

Thus, the expansion of $(x+a)^{m+1}$ is exactly of the same form as that of $(x+a)^m$, Hence if the theorem is true for the index m, it is also true for the index m+1.

Step-III: But we have seen that the theorem is true for the value n = 3. Therefore it should be true for the value 3+1 =4. Hence the theorem is true for all positive integral values of n.

Remarks:

1. $(x-a)^n = x^n + {}^nC_1(-a)x^{n-1} + {}^nC_2(-a)^2x^{n-2} + {}^nC_3(-a)^3x^{n-3} + \ldots\ldots + {}^nC_n\,(-a)^n$

$$= x^n - {}^nC_1(a)x^{n-1} + {}^nC_2\,(a)^2x^{n-2} - {}^nC_3\,(a)^3x^{n-3} + \ldots\ldots + (-1)^n\,{}^nC_n\,(a)^n$$

Thus, the various terms in the expansion of $(x-a)^n$ are numerically same as those of $(x+a)^n$ with the difference that here the terms are alternatively positive and negative and the last term is positive or negative depending on whether 'n' is even or odd respectively.

3. By interchanging x and a in (1.6) we have,

$$(a+x)^n = a^n + {}^nC_1 a^{n-1} x + {}^nC_2 a^{n-2} x^2 + \ldots + {}^nC_n x^n$$

Putting a = 1 we get,

$$(1+x)^n = 1 + {}^nC_1 x + {}^nC_2 \; x^2 + \ldots + {}^nC_n \; x^n$$

a formula convenient to remember.

1.6.2. General term: In the expansion of $(x+a)^n$, we find that the

coefficient of the 1st term = 1 = ${}^nC_0 = {}^nC_{1-1}$

coefficient of the 2nd term = ${}^nC_1 = {}^nC_{2-1}$

coefficient of the 3rd term = ${}^nC_2 = {}^nC_{3-1}$

So, the coefficient of the (r+1)th term = ${}^nC_{r+1-1} = {}^nC_r$

Denoting the (r+1)th term by t_{r+1}, we have

$$t_{r+1} = {}^nC_r \, x^{n-r} a^r = \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!} x^{n-r} a^r$$

t_{r+1} is called the general term because, any required term can be obtained from it by giving a suitable value to r. For example, to obtain the 5th term from the beginning, putting r = 4 in t_{r+1} we get $t_5 = {}^nC_4 \; x^{n-4} a^4$. The (r+1)th term is also denoted by T_{r+1}.

Note that in the general term the index of x is (n - r) and that of 'a' is 'r'. So the general term in the expansion of $(x-a)^n$ is ${}^nC_r \, x^{n-r} (-a)^r = (-1)^r \; {}^nC_r \, x^{n-r} a^r$

a) The simplest form of the binomial expansion is,

$$t_{r+1} = \frac{n(n-1)(n-2)....(n-r+1)}{r!}x^r$$

b) $$(x+y)^n = \left[x(1+\frac{y}{x})\right]^n = x^n(1+z)^n, \text{ where } z = \frac{y}{x}$$

$$= x^n\,(1 + {}^nC_1\, z + {}^nC_2\, z^2 + ...+ {}^nC_r\, z^r + ...+ {}^nC_n z^n)$$

$$= x^n + {}^nC_1\, x^n . \frac{y}{x} + {}^nC_2\, x^n . \frac{y^2}{x^2} + ...+ {}^nC_r\, x^n . \frac{y^r}{x^r} + ...+ {}^nC_n\, x^n . \frac{y^n}{x^n}$$

$$= x^n + {}^nC_1\, x^{n-1}\, y + {}^nC_2\, x^{n-2}\, y^2 + ...+ {}^nC_r\, x^{n-r}\, y^r + ...+ {}^nC_n\, y^n$$

Thus, the expansion of $(x+y)^n$ can be obtained from that of $(1+z)^n$

c) The binomial expansion of $(1-x)^n$ is obtained by writing $(-x)$ in place of x in (1.4). Thus,

$$(1-x)^n = 1 - nx + \frac{n(n-1)}{12}x^2 - \frac{n(n-1)(n-2)}{123}x^3 + ...+(-1)^n x^n$$

e.g. $(1-x)^5 = 1 - 5x + 10x^2 - 10x^3 + 5x^4 - x^5$

The general term in the above expansion is given by

$$t_{r+1} = (-1)^r\ \frac{n(n-1)(n-2)....(n-r+1)}{r!}x^r$$

d) The coefficient of the first term in the expansion of $(x+a)^n$ is 1 or ${}^nC_0$. The coefficients of second, third, fourth ..... and (n+1)th terms are ${}^nC_1$, ${}^nC_2$, ${}^nC_3$, ... and ${}^nC_n$ respectively. These coefficients are called binomial coefficients. Binomial coefficients are some times denoted by C_0, C_1, C_2, ...C_n.

2!   3!

e) The coefficient of (r+1)th term in the expansion of $(x+a)^n$ is the sum of the coefficients of rth and (r+1)th terms in the expansion of $(x+a)^{n-1}$. For example, as the coefficients of $(x+a)^2$ are 1,2 and 1, the coefficients of $(x+a)^3$ would be 1, (1+2 = 3), (2 +1 = 3), 1. The coefficients of $(x+a)^4$ would be 1, (1+3 = 4), (3 +3 = 6), (3 +1 = 4), 1. etc. So, when the coefficients in the expansion of $(x+a)^k$ are known, the coefficients in the expansion of $(x+a)^{k+1}$ can be obtained without doing the actual expansion.

Example: 1.27 How many terms are there in the expansion of $[(2x+3y)^2]^5$?

Solution: $[(2x+3y)^2]^5 = (2x+3y)^{10}$

Hence, the index, n = 10 is a positive integer.

So, the number of terms in the expansion = n +1 = 10 +1 = 11.

Example: 1.28 Write down the expansion of $\left(3x-\frac{1}{2}y\right)^4$ by the Binomial Theorem. By giving suitable values to x and y, obtain the value of $(29.5)^4$ correct to five significant figures.

Solution:

$$\left(3x-\frac{1}{2}y\right)^4 = (3x)^4 + {}^4C_1(3x)^3\left(-\frac{y}{2}\right) + {}^4C_2(3x)^2\left(-\frac{y}{2}\right)^2 + {}^4C_3(3x)\left(-\frac{y}{2}\right)^3 + \left(-\frac{y}{2}\right)^4$$

$$= (3x)^4 - 4(3x)^3\left(\frac{y}{2}\right) + 6(3x)^2.\frac{y^2}{4} - 4(3x)^2\frac{y^3}{8} + \frac{y^4}{16} \qquad \text{......(1)}$$

$$= 81x^4 - 54x^3y + \frac{27}{2}x^2y^2 - \frac{3}{2}xy^3 + \frac{1}{16}y^4$$

Example: 1.29 Expand and simplify

$$\left(\sqrt{2}+1\right)^6+\left(\sqrt{2}-1\right)^6$$

Solution:

$$\left(\sqrt{2}+1\right)^6=\left(\sqrt{2}\right)^6+{}^6C_1\left(\sqrt{2}\right)^5.1+{}^6C_2\left(\sqrt{2}\right)^4.1^2+{}^6C_3\left(\sqrt{2}\right)^3.1^3$$

$$+{}^6C_4\left(\sqrt{2}\right)^2.1^4+{}^6C_5\left(\sqrt{2}\right).1^5+{}^6C_6.1^6$$

$$\left(\sqrt{2}-1\right)^6=\left(\sqrt{2}\right)^6-{}^6C_1\left(\sqrt{2}\right)^5.1+{}^6C_2\left(\sqrt{2}\right)^4.1^2-{}^6C_3\left(\sqrt{2}\right)^3.1^3$$

$$+{}^6C_4\left(\sqrt{2}\right)^2.1^4-{}^6C_5\left(\sqrt{2}\right).1^5+{}^6C_6.1^6$$

Adding, we have, $\left(\sqrt{2}+1\right)^6+\left(\sqrt{2}-1\right)^6=2\left\{\left(\sqrt{2}\right)^6+{}^6C_2\left(\sqrt{2}\right)^4+{}^6C_4\left(\sqrt{2}\right)^2+{}^6C_6\right\}$

(terms with odd powers of $\left(\sqrt{2}\right)$ cancel out)

$$=2\left\{2^3+\frac{6\times5}{1.2}.2^2+\frac{6\times5\times4\times3}{4.3.2.1}.2+1\right\}$$

$$=2(8+60+30+1)=198$$

Example: 1.30 Find the 10th term of $\left(2x^2+\frac{1}{x}\right)^{12}$

Solution: Here n =12, 'x' = $2x^2$, 'a' = $\frac{1}{x}$

$$T_{r+1}= {}^nC_r\, x^{n-r}\, a^r= {}^{12}C_r\ (2x^2)^{12-r}\left(\frac{1}{x}\right)^r$$

Example: 1.31 Which term contains x^8 in the expansion of $\left(x^2-\frac{1}{x}\right)$? Also find its coefficient.

Solution: Here n =10, 'x' = x^2 and 'a' = $-\frac{1}{x}$.

Let the (r+1)th term contains x^8.

$$T_{r+1} = {}^{10}C_r\ (x^2)^{10-r}\left(-\frac{1}{x}\right)^r = {}^{10}C_r\ x^{20-2r}\ \frac{(-1)^r}{x^r}$$

$$= (-1)^r\ .\ {}^{10}C_r\ x^{20-3r} \quad \text{.......} (1)$$

To find the term containing x^8 , we take $x^{20-3r} = x^8$

i.e. $20-3r = 8$ or, $-3r = -12$ or, $r = 4$

Putting r = 4 in (1), we get

$$T_5 = (-1)^4.{}^{10}C_4.x^8 = \frac{10\times9\times8\times7}{4\times3\times2\times1}\times x^8 = 210x^8$$

So, the 5th term contains x^8 and its coefficient is 210.

Example: 1.32 Write and simplify the term involving x^{18} in $\left(ay^{\frac{1}{2}}-bx^3\right)^{17}$.

Solution: In the expansion of $\left(ay^{\frac{1}{2}}-bx^3\right)^{17}$, the (r+1)th term is:

$$T_{r+1} = {}^{17}C_r\left(ay^{\frac{1}{2}}\right)^{(17-r)}\ (-bx^3)^r$$

$$T_{6+1} = T_7 = {}^{17}C_6\, a^{11} y^{\frac{11}{2}} (-b)^6 x^{18}$$

$$= 12376\ a^{11} b^6 y^{11/2} x^{18}$$

Example: 1.33 Write the general term in the expansion of $(x^2 - y)^6$

Solution: Let T_{r+1} be the general term in the expansion of $(x^2 - y)^6$

$$T_{r+1} = {}^6C_r\ (x^2)^{6-r}\,(-y)^r = (-1)^r\ {}^6C_r\, x^{12-2r}\, y^r$$

Example: 1.34 Find the 5th term from the end in the expansion of $\left(\frac{x^3}{2} - \frac{2}{x^3}\right)^9$

Solution: Here the index $n = 9$

Total number of terms in the expansion is $9 + 1 = 10$

The 5th term from the end has $(10 - 5) = 5$ terms before it and is the 6th term from the beginning.

$$\text{Hence } T_{r+1} = {}^nC_r\, x^{n-r} a^r = {}^9C_r \left(\frac{x^3}{2}\right)^{9-r} \times \left(-\frac{2}{x^3}\right)^r$$

Putting $r = 5$,

$$T_6 = {}^9C_5 \left(\frac{x^3}{2}\right)^4 \times \left(-\frac{2}{x^3}\right)^5 = \frac{9!}{5!4!} \times \frac{x^{12}}{2^4} \times -\frac{2^5}{x^{15}}$$

$$= \frac{9 \times 8 \times 7 \times 6}{4 \times 3 \times 2 \times 1} \times x^{12-15} \times (-2) = -3 \times 7 \times 6 \times 2 \times x^{-3} = -252 / x^3$$

1.6.3. Middle Terms:

In the binomial expansion of $(x+a)^n$ or $(1+x)^n$ there would be either one middle term or two middle terms depending on the value of 'n' as even or odd.

i.e., ${}^{n}C_{\frac{n}{2}}\, x^{\frac{n}{2}}\, a^{\frac{n}{2}}$ and is equal to $\dfrac{n!}{\left(\frac{n!}{2}\right)^2} x^{\frac{n}{2}}\, a^{\frac{n}{2}}$.

Case II : When 'n' is odd.

If 'n' is odd and is equal to (2m+1), the total number of terms being (2m+2) is even. In this case there are two middle terms viz. the (m+1)th and the (m+2)th terms. Since $2m+1 = n$, $m = \dfrac{n-1}{2}$. So, the middle terms are $\left(\dfrac{n-1}{2}+1\right)$th and $\left(\dfrac{n+1}{2}+1\right)$th terms and are ${}^{n}C_{\frac{n-1}{2}}\, x^{\frac{n+1}{2}}\, a^{\frac{n-1}{2}}$ and ${}^{n}C_{\frac{n+1}{2}}\, x^{\frac{n-1}{2}}\, a^{\frac{n+1}{2}}$ .On simplification, these become

$$\frac{n!}{\left(\frac{n-1}{2}\right)!\left(\frac{n+1}{2}\right)!} x^{\frac{n+1}{2}} . a^{\frac{n-1}{2}} \text{ and } \frac{n!}{\left(\frac{n+1}{2}\right)!\left(\frac{n-1}{2}\right)!} x^{\frac{n-1}{2}} . a^{\frac{n+1}{2}}$$

Thus, the numerical values of the coefficients of the two middle terms are equal.

If n = 8, the number of terms is 8 +1 = 9.

So the 5th term is the middle term

If n = 9, the number of terms is 9+1 = 10

So there are two middle terms, viz. the fifth and the sixth terms i.e.

$\left(\dfrac{9-1}{2}+1\right)$ th and $\left(\dfrac{9+1}{2}+1\right)$ th i.e. the 5 th and the 6 th terms.

So, $t_r = {}^nC_{r-1}.x \;.a$ and $t_{r+1} = {}^nC_r x \; a$

Now, $\dfrac{t_{r+1}}{t_r} = \dfrac{{}^nC_r}{{}^nC_{r-1}}.\dfrac{a}{x} = \dfrac{n-r+1}{r}.\dfrac{a}{x}$

$$t_{r+1} \geq t_r \text{ if } \frac{t_{r+1}}{t_r} \geq 1 \text{ i.e., if } \frac{n-r+1}{r}.\frac{a}{x} \geq 1$$

i.e., if $(n-r+1)\, a \geq rx$

i.e., if $(an-ar+a) \geq rx$

i.e., if $(n+1)a \geq r(a+x)$

i.e., if $r \leq \dfrac{(n+1)a}{a+x} = k$, say

Thus the terms continue to increase as long as $r < k$. When 'k' is an integer, $t_{k+1} = t_k$ and these are the two greatest terms.

If $\dfrac{(n+1)a}{x+a}$ is a fraction, let its integral part be denoted by m, i.e., $\dfrac{(n+1)a}{x+a} = (m+f)$, where 'm' is the integral part and 'f' is the fractional part.

The terms continue to increase till $r = m$. Thus t_{m+1} is the greatest term in the expansion of $(x+a)^n$.

1.6.5. Equidistant Terms:

In the expansion of $(a+x)^n$ or $(1+x)^n$, the coefficients of terms equidistant from the beginning and the end are equal.

is equal to the coefficient of the (r+1)th term from the end.

It can be verified that, $^nC_0 = {}^nC_n$, $^nC_1 = {}^nC_{n-1}$, $^nC_2 = {}^nC_{n-2}$ etc.

Example : If $x = \frac{1}{3}$, find the greatest term in the expansion of $(1+4x)^8$ without assuming the formula. Also compute its value.

Solution: Let T_{r+1} be the greatest term in the expansion of $(1+4x)^8$.

Now, $T_{r+1} = {}^8C_r(4x)^r$ and $T_r = {}^8C_{r-1}(4x)^{r-1}$

$$\therefore \quad \frac{T_{r+1}}{T_r} = \frac{8-r+1}{r}.4x$$

Using the given value $x = \frac{1}{3}$, we have,

$$\frac{T_{r+1}}{T_r} = \frac{8-r+1}{r}.4.\frac{1}{3} = \frac{9-r}{r}.\frac{4}{3} = \frac{36-4r}{3r}$$

If T_{r+1} is the greatest term, then $T_{r+1} > T_r$ or, $\frac{T_{r+1}}{T_r} > 1$, or, $\frac{36-4r}{3r} > 1$ or $7r < 36$.

Hence $r < 5\frac{1}{7}$ i.e., r = 5

Hence, $T_{r+1} = T_6$ is the greatest term.

Now, $T_6 = {}^8C_5(4x)^5 = {}^8C_5\left(4.\frac{1}{3}\right)^5 = \frac{8.7.6}{1.2.3}\left(\frac{4}{3}\right)^5 = 56\left(\frac{4}{3}\right)^5$

= 57344/243

Thus, $T_6 = {}^{10}C_5 (x)^{10-5}\left(-\frac{1}{2y}\right)^5$ $\left[\text{Since } T_{r+1} = {}^{n}C_r x^{n-r} a^r \text{ in } (x + a)^n\right]$

$$= {}^{10}C_5\, x^5\left(-\frac{1}{2y}\right)^5$$

$$= \frac{10 \times 9 \times 8 \times 7 \times 6}{5 \times 4 \times 3 \times 2 \times 1} \cdot \frac{x^5(-1)^5}{2^5 y^5} = 252\left(\frac{-x^5}{2^5 y^5}\right) = -\frac{63}{8}\frac{x^5}{y^5}$$

Example: 1.37 Find the middle terms in the expansion of $\left(3x - \frac{x^3}{6}\right)^7$

Solution: Here n = 7. So, the number of terms in the expansion = 7 +1 = 8

There are two middle terms viz $\left(\frac{7-1}{2}+1\right)$th and $\left(\frac{7+1}{2}+1\right)$th terms which are the 4th and the 5th terms.

So. $T_4 = {}^{7}C_3 (3x)^{7-3}\left(-\frac{x^3}{6}\right)^3 = \frac{7 \times 6 \times 5}{3 \times 2 \times 1}(3x)^4\left(-\frac{x^9}{216}\right)$

$$= 35.3^4.x^4\left(-\frac{x^9}{216}\right) = -\frac{35 \times 81}{216}x^{13} = -\frac{105}{8}x^{13}$$

Example: 1.38 Find the term independent of x in the expansion of $\left(x-\frac{1}{x}\right)^{16}$

Solution: Let t_{r+1} be the term independent of x in the expansion of $\left(x-\frac{1}{x}\right)^{16}$

Now $t_{r+1}={}^{16}C_r(x)^{16-r}\left(-\frac{1}{x}\right)^r={}^{16}C_r x^{16-r}(-1)^r.x^{-r}$

$$=(-1)^r\ {}^{16}C_r x^{16-2r}$$

If this term is independent of x, then the index of x should be zero.

So, $16-2r=0$ or $r=8$

Hence t_9 is the term independent of x.

So, $t_9={}^{16}C_8(-1)^8x^{16-16}$

$$=\frac{16!}{8!8!}=12870$$

Example: 1.39 Use Binomial Theorem to evaluate $(999)^4$

Solution: $(999)^4=(1000-1)^4$

$$={}^4C_0\ (1000)^4-{}^4C_1\,(1000)^3+{}^4C_2\ (1000)^2-{}^4C_3\ (1000)+{}^4C_4\ (-1)^4$$

$$=(1000)^4-4\,(1000)^3+6(1000)^2-4(1000)+1$$

$$=996005996001$$

2. (a) Evaluate: $\left[x+\sqrt{x^2-1}\right] + \left[x-\sqrt{x^2-1}\right]$

(b) Find: $\left(\sqrt{x}+\sqrt{y}\right)^6 + \left(\sqrt{x}-\sqrt{y}\right)^6$

(c) Find: $\left(\sqrt{2}+1\right)^5 + \left(\sqrt{2}-1\right)^5$

3. Use the Binomial Theorem to evaluate

(i) $(99)^4$ (ii) $(102)^6$

4. (a) Write down the 7th term in the expansion of $\left(\frac{4x}{5}-\frac{5}{2x}\right)^9$

(b) Write the 10th term in the expansion of $(x - y^2)^{15}$.

(c) Find the term containing x^2, if any, in the expansion of $\left(3x-\frac{1}{2x}\right)^8$

(d) Find the coefficient of x^{15} in the expansion of $(x^3 - 2)^{11}$.

5. (a) Find the middle term in the expansion of $\left(\frac{a}{x}-bx\right)^{12}$

(b) Find the two middle terms in the expansion of $\left(3x-\frac{2x^2}{3}\right)^7$

6. Show that the middle term in the expansion of $(1+x)^{2n}$ is $\frac{1.3.5....(2n-1)}{n!}.2^n.x^n$.

7. Prove that the middle term of $\left(x+\frac{1}{2x}\right)^{2n}$ is $\frac{1.3.5....(2n-1)}{n!}$.

10. Find the value of $(1.0005)^4$ to four decimal places by Binomial Theorem.

11. In the binomial expansion of $(1+a)^{m+n}$, prove that the coefficients of a^m and a^n are equal.

12. (a) Prove that the coefficient of x^n in $(1+x)^{2n}$ is twice the coefficient of x^n in $(1+x)^{2n-1}$.

(b) Find the coefficient of x in $(1-2x+3x^2)(1-x)^{14}$.

13. If x^p occurs in the expansion of $\left(x^2+\frac{1}{x}\right)^{2n}$, prove that its coefficient is

$$(2n)!/\left(\frac{4n-p}{3}\right)!\left(\frac{2n+p}{3}\right)!$$

14. Prove that the coefficient of the (r+1)th term in the expansion of $(1+x)^{n+1}$ is equal to the sum of the coefficients of the rth and (r+1)th terms of $(1+x)^n$.

15. Show that the coefficient of the middle term in the expansion of $(1+x)^{2n}$ is equal to the sum of the coefficients of the two middle terms in the expansion of $(1+x)^{2n-1}$.

16. (a) In the expansion of $(1+x)^{43}$, if the coefficients of (2r+1)th and (r+2)th terms are equal, find r.

(b) In the binomial expansion of $(a+b)^n$, the coefficients of the fourth and the thirteenth terms are equal. Find n.

17. The coefficients of (r – 1)th, rth and (r+1)th terms in the expansion of $(x+1)^n$ are in the ratio 1:3:5. Find, both n and r.

18. If T_r is the rth term in the expansion of $(1+a)^n$ in ascending powers of a, prove that $r(r+1)T_{r+2}=(n-r+1)(n-r)a^2T_r$.

expression is given by $1 + a_1x + a_2x^2 + ... + a_{12}x^{12}$, show that, $a_2 + a_4 + a_6 + ...+ a_{12} = 31$.

22. In the expansion of $(1+x)^{10}$, the coefficient of (2r+1)th term is equal to the coefficient of (4r +5)th term. Find r.

23. If x^r occurs in the expansion of $\left(x+\frac{1}{x^2}\right)^{2n}$, prove that its coefficient is

$$\frac{(2n)!}{\left\{\frac{1}{3}(2n-r)\right\}!\left\{\frac{1}{3}(4n+r)\right\}!}$$

24. (a) If the coefficient of x in the expansion of $\left(x^2+\frac{k}{x}\right)^5$ is 270, find k.

(b) If the absolute term in the expansion of $\left(\sqrt{x}-\frac{K}{x^2}\right)^{10}$ is 405, find the value of K.

25. (a) If the 21st and 22nd terms in the expansion of $(1+x)^{44}$ are equal , find the value of x.

(b) In the expansion of $(1+x)^{11}$, the fifth term is 24 times the third term. Find the value of x.

26. If a_1, a_2, a_3, a_4 are the coefficients of the second, third, fourth and fifth terms respectively in the binomial expansion $(1+x)^n$, prove that

$$\frac{a_1}{a_1+a_2}+\frac{a_3}{a_3+a_4}=\frac{2a_2}{a_2+a_3}$$

***

Proof: Let $(1+x)^n = C_0 + C_1 x + C_2 x^2 + \ldots\ldots + C_n x^n$ (1.5)

Putting x = 1 on both sides of (1.5), we have

$$(1+1)^n = C_0 + C_1 + C_2 + \ldots\ldots + C_n$$

$$\text{i.e. } C_0 + C_1 + C_2 + \ldots\ldots + C_n = 2^n \quad (1.6)$$

Corollary 1:

Total number of combinations of 'n' different objects taken some or all at a time can be found by using the equation (1.6).

Thus, we have, $1 + C_1 + C_2 + \ldots\ldots + C_n = 2^n$

So, $C_1 + C_2 + \ldots\ldots + C_n = 2^n - 1$

This shows that the total number of combinations of 'n' different objects taken 1 or 2 or ...or n at a time is $2^n - 1$.

Corollary 2: The number of all subsets of a finite set of 'n' elements is 2^n.

Proof: We know that an empty set is a set containing no element and is a subset of every set. There is only 1 = C(n,0) = C_0 such set with no element. We also know that the number of combinations of 'n' distinct objects taken 1 at a time is C(n,1). Thus the number of 1 element subsets of 'n' distinct objects = C(n,1) = C_1. Similarly , the number of 2 element subsets is C(n,2) = C_2; that of 3 element subsets is C(n,3) = C_3 and so on. Finally, the number of 'n' element subsets is C(n, n) = C_n.

The total number of subsets, therefore, is

$$C_0 + C_1 + C_2 + \ldots. + C_n = 2^n \quad \text{(by property I)}$$

Hence, the number of all subsets of a finite set with 'n' elements is 2^n.

⇒ Sum of the odd coefficients = Sum of the even coefficients

Since $C_0 + C_1 + C_2 + ... + C_n = 2^n$

$$C_0 + C_2 + C_4 + ... = C_1 + C_3 + C_5 + ... = \frac{1}{2} . 2^n = 2^{n-1}.$$

Property III: In the expansion of $(1+x)^n$, where 'n' is a positive integer, coefficients of terms equidistant from the beginning and the end are equal.

For the proof, see 1.6.5

Example: 1.40 The sum of the squares of the coefficients in the expansion of $(1+x)^n$ is $(2n)!/(n!)^2$

Solution: We know that $(1+x)^n = C_0 + C_1x + C_2x^2 + ...+C_{n-1} x^{n-1} + C_n x^n$ (1)

Also, $(x+1)^n = C_0 x^n + C_1x^{n-1} + C_2 x^{n-2} + ... +C_{n-1} x + C_n$ (2)

Multiplying (1) and (2), we get

$$(1+x)^{2n} = (C_0 + C_1 x + C_2x^2 + ...+C_{n-1} x^{n-1} + C_n x^n) \text{ x}$$

$$(C_0 x^n + C_1 x^{n-1} + C_2 x^{n-2} + ...+C_{n-1} x + C_n) \quad (3)$$

Equating the coefficients of x^n on both sides, we get

$$^{2n}C_n = C_0^{\,2} + C_1^{\,2} + C_2^{\,2} + ... + C_{n-1}^{\,2} + C_n^{\,2}$$

So, the sum of the squares of the binomial coefficients is

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

Remarks: Equating coefficients of x^{n+1} on both sides of (3), we get

$$^{2n}C_{n+1} = C_0 C_1 + C_1 C_2 + C_2 C_3 + ... + C_{n-1} C_n$$

(ii) $^{15}C_2+^{15}C_3+^{15}C_4+......+^{15}C_{15}$

Solution:

(i) We know that, $^nC_0+^nC_1+^nC_2+....+^nC_n = 2^n$ ..... (I)

Putting n = 11 in (I), we get,

$$^{11}C_0+^{11}C_1+^{11}C_2+.....+^{11}C_{11} = 2^{11}$$

(ii) $^{15}C_2+^{15}C_3+^{15}C_4+.....+^{15}C_{15}$

$$=(^{15}C_0+^{15}C_1+^{15}C_2+^{15}C_3+.....+^{15}C_{15})-(^{15}C_0+^{15}C_1)$$

$$=2^{15}-(1+15)=2^{15}-16 = 2^4(2^{11}-1)=32768$$

Example: 1.42

Using Binomial Theorem, prove that 6^n - 5n always leaves the remainder 1 when divided by 25.

Solution:

By Binomial Theorem, we have

$$6^n=(1+5)^n=1+\ ^nC_1.5+^nC_2.5^2+^nC_3.5^3+....+^nC_n.5^n$$

$$=1+5n+5^2\ ^nC_2+5^3\ ^nC_3+....+5^n$$

$$\therefore\ 6^n-5n=5^2\left[^nC_2+5.^nC_3+....+5^{n-2}\right]+1$$

= 25 (a positive integer) +1

$\therefore$ $(6^n-5n)\div 25$ = a positive number (quotient) + 1 (remainder).

Hence, 6^n-5n always leaves the remainder 1 when divided by 25.

Example: 1.43

If $C_0, C_1, C_2, ..., C_n$ are binomial coefficients in the expansion of $(1+x)^n$,

Solution:

(i) $\text{L.H.S} = C_1 + 2.C_2 + 3.C_3 + ... + n.C_n$

$$= n + 2.\frac{n(n-1)}{2.1} + 3.\frac{n(n-1)(n-2)}{3.2.1} + ... + n.1$$

$$= n\left[1 + (n-1) + \frac{(n-1)(n-2)}{2.1} + ... + 1\right]$$

$$= n\left[{}^{n-1}C_0 + {}^{n-1}C_1 + {}^{n-1}C_2 + ... + {}^{n-1}C_{n-1}\right] = n(1+1)^{n-1} = n.2^{n-1} = \text{R.H.S}$$

$$(\text{since} \quad C_0 + C_1 + C_2 + ... + C_n = 2^n)$$

(ii) $\text{L.H.S} = C_0 + 2C_1 + 3C_2 + ... + (n+1)\,C_n$

$$= C_0 + (C_1 + C_1) + (C_2 + 2C_2) + ... + (C_n + nC_n)$$

$$= (C_0 + C_1 + C_2 + ... + C_n) + (C_1 + 2C_2 + ... + n\,C_n)$$

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

(iii) $\text{LHS} = C_0 + 3C_1 + 5C_2 + ... + (2n+1)\,C_n$

$$= C_0 + (C_1 + 2C_1) + (C_2 + 4C_2) + ... + (C_n + 2nC_n)$$

$$= (C_0 + C_1 + C_2 + ... + C_n) + 2(C_1 + 2C_2 + 2C_3 .... + n\,C_n)$$

$$= 2^n + 2.n.2^{n-1}$$

$$= 2^n + n.\,2^n = (n+1)\,2^n = \text{R.H.S.}$$

Example:1.44

Prove that

(i) $C_0 + 2C_1 + 2^2C_2 + ... + 2^nC_n = 3^n$

Or, $C_0 + 2C_1 + 2^2C_2 + ... + 2^nC_n = 3^n$

(ii) $C_0 - 2C_1 + 3C_2 + ... + (-1)^n (n+1)C_n$

$= [C_0 - C_1 + C_2 - C_3 + ... + (-1)^n C_n] - [C_1 - 2C_2 + 3C_3 + ... - (-1)^{n-1} nC_n]$

$= (1-1)^n - (C_1 - 2C_2 -+ 3C_3 - + ... + (-1)^{n-1} nC_n)$

$$= 0 - \left[n - 2.\frac{n(n-1)}{2!} + 3.\frac{n(n-1)(n-2)}{3.\ 2.\ 1} .... + (-1)^{n-1}.n.1\right]$$

$$= -n\left[1 - (n-1) + \frac{(n-1)(n-2)}{2.1} .... + (-1)^{n-1}\right]$$

$$= -n\left[{}^{n-1}C_0 - {}^{n-1}C_1 + {}^{n-1}C_2 .... + (-1)^{n-1}\ {}^{n-1}C_{n-1}\right]$$

$= -n\ (1-1)^{n-1} = -n.\ 0 = 0$

EXERCISES- 1.3

1. Find the value of

(i) ${}^{12}C_0 + {}^{12}C_1 + {}^{12}C_2 + ... + {}^{12}C_{12}$

(ii) ${}^{15}C_1 + {}^{15}C_2 + {}^{15}C_3 + ... + {}^{15}C_{15}$

2. Write down the binomial expansion of $(1+x)^{n+1}$ when $x = 8$. Deduce that $9^{n+1} - 8n - 9$ is divisible by 64, whenever 'n' is a positive integer.

3. If $(1+x)^n = C_0 + C_1 x + C_2 x^2 + ... + C_n x^n$, prove that

(a) $$\frac{C_1}{C_0} + 2\frac{C_2}{C_1} + 3\frac{C_3}{C_2} + ... + n\frac{C_n}{C_{n-1}} = \frac{n(n+1)}{2}$$

(d) $(C_0 + C_1)(C_1 + C_2)(C_2 + C_3)\ldots(C_{n-1} + C_n) = \frac{\quad}{n!}$

4. If $C_0, C_1, C_2, \ldots C_n$ are the coefficients of successive terms in the expansion of $(1+x)^n$, where 'n' is a positive integer, prove that

(i) $C_0 + \frac{C_1}{2} + \frac{C_2}{3} + \ldots + \frac{C_n}{n+1} = \frac{2^{n+1} - 1}{n+1}$

(ii) $C_0 + \frac{1}{3}C_2 + \frac{1}{5}C_4 + \ldots = \frac{2^n}{n+1}$

(iii) $C_0 - \frac{C_1}{2} + \frac{C_2}{3} + \ldots + (-1)^n \frac{C_n}{n+1} = \frac{1}{n+1}$

5. If $C_0, C_1, C_2, \ldots C_n$ be the coefficients in the expansion of $(1+x)^n$ where 'n' is a positive integer, prove that

$$C_0C_1 + C_1C_2 + C_2C_3 + \ldots + C_{n-1}C_n = \frac{2^n . n[1\ 3.\ 5 \ldots (2n-1)]}{(n+1)!}$$

6. If $C_0, C_1, C_2, \ldots C_n$ denote the coefficients in the expansion of $(1+x)^n$, prove that

$$C_0C_r + C_1C_{r+1} + C_2C_{r+2} + \ldots + C_{n-r}C_n = \frac{(2n)!}{(n-r)!(n+r)!}$$

7. The coefficients of the 5th, 6th and 7th terms in the expansion of $(1+x)^n$ are in A.P. Find 'n'.

8. The coefficients of three consecutive terms in the expansion of $(1+x)^n$ are in the ratio 1:7:42. Find 'n'.

9. A particular three consecutive coefficients in the expansion of $(1+x)^n$ are in the ratio 1:3:5. Find 'n'.

$1, -4x, 7x^2$. Find the values of 'a' and 'n'.

12. Fill in the blanks.

(i) The middle term in the expansion $\left(\frac{2x^2}{3}+\frac{3}{2x^2}\right)^{10}$ is ______

(ii) Let $(1+x)^n = C_0 + C_1x + C_2x^2 + \ldots + C_nx^n$

If $\frac{C_1}{C_0}+\frac{2C_2}{C_1}+\frac{3C_3}{C_2}+\ldots+\frac{nC_n}{C_{n-1}}=\frac{1}{k}.n(n+1)$, then k = ______

(iii) The larger of $(99^{50}+100^{50})$ and 101^{50} is ______

(iv) The sum of the coefficients of the polynomial $(1+x+x^2)^{143}$ is ______

13. Some statements are given below. If the statement is correct, write True otherwise write False.

(i) The term independent of x in the expansion of $\left(2x^2-\frac{1}{x}\right)^{12}$ is 6920.

(ii) If 'P' be the sum of the odd terms and 'Q' that of the even terms in the expansion of $(x+a)^n$, then $(x^2-a^2)^n = P^2-Q^2$

(iii) $C_0^2 + C_1^2 + C_2^2 + \ldots + C_n^2 = {}^{2n}C_n$

14. Choose the correct alternative

(i) The coefficient of x^{32} in the expansion of $\left(x^4-\frac{1}{x^2}\right)^{15}$ is

(a) 1165 (b) 1265 (c) 1365 (d) None of these

(iv) ${}^9C_0+{}^9C_2+{}^9C_4+{}^9C_6+{}^9C_8$ is equal to

(a) 255 (b) 256 (c) 257 (d) None of these

***

1.6.7.Binomial Theorem for any index:-

We have shown that $(1+x)^n$ can be expanded in ascending powers of x as $(1+x)^n = C(n,0) + C(n,1)x + ...+ C(n,r)x^r + ...+ C(n,n)x^n$.

This expansion is valid for all values of x only if n, the index, is a positive integer. But when 'n' is negative or a fractional, the symbols C(n,0), C(n,1), C(n,2), etc are meaningless and so cannot be computed. However, the same can be done by writing $(1+x)^n$ in the following form

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + ... + \frac{n(n-1)....(n-r+1)}{r!}x^r + ... \qquad (1.7)$$

We make the following observations:

For a negative or fractional value of 'n'.

(i) equation (1.7) is valid only if the value of x is numerically less than 1 i.e., $|x| < 1$ or $-1 < x < 1$.

(ii) the number of terms on the RHS of (1.7) is infinite.

We have already given a proof of (1.7) for positive integral values of 'n'. For negative or fractional values of 'n', the proof is complicated and is beyond the scope of the present text. We simply assume that (1.7) is valid for negative or fractional values of 'n' provided $|x|<1$.

Note

(1): The expansion of $(a+b)^n$ for negative or fractional values of 'n' canbe found as follows:

$$= a^n + na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 + ... + \frac{n(n-1)..(n-r+1)}{r!}a^{n-r}b^r + ... \qquad (1.8)$$

(2): We consider the following example to illustrate that for negative value of the index 'n', the expansion

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + ... + \frac{n(n-1)....(n-r+1)}{r!}x^r + ...$$

Is not valid when $|x| \geq 1$.

Let $n = -1$, and $x = -2$

Putting these values in equation (1.7) above, we have,

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

$$+.....+ \frac{(-1)(-1-1)(-1-2)....(-1-r+1)}{r!}(-2)^r + ...$$

$$= 1 + 2 + \frac{1.2}{2!}.2^2 + \frac{1.2.3}{3!}2^3 + ... + \frac{1.2....r}{r!}2^r$$

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

But the LHS of (1.7) is $(1-2)^{-1} = (-1)^{-1} = \frac{1}{-1} = -1$

Such an absurd result is obtained by taking $x = -2$ in the expansion of $(1+x)^{-1}$

Students are advised to remember the following facts:

1. For negative or fractional values of the index, the number of terms in the expansion of $(1+x)^n$ is infinite because $(n-r+1)$ will never vanish and as such, there is no last term. In other words, the expansion continues without end.

that, $|\frac{x}{a}| < 1$

1.6.8: General Term:

The (r+1)th term , t_{r+1} is called the general term and is given by

$$t_{r+1} = \frac{n(n-1)(n-2)....(n-r+1)}{r!}x^r$$

Cor 1: Expansion of $(1+x)^{-n}$, $|x| < 1$

We know that

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + ... + \frac{n(n-1)(n-2)....(n-r+1)}{r!}x^r + ... \text{to } \infty \qquad (1.9)$$

Changing n to $-n$ in (1.9), we get

$$(1+x)^{-n} = 1 + (-n)x + \frac{(-n)(-n-1)}{2!}x^2 + ... + \frac{(-n)(-n-1)(-n-2)....(-n-r+1)}{r!}x^r + ... \text{to } \infty$$

$$= 1 - nx + \frac{n(n+1)}{2!}x^2 - .... + (-1)^r \frac{n(n+1)(n+2)....(n+r-1)}{r!}x^r + ... \text{to } \infty \qquad (1.10)$$

Thus the general term in $(1+x)^{-n}$ is $(-1)^r \frac{n(n+1)(n+2)....(n+r-1)}{r!}x^r$

Cor 2: Expansion of $(1-x)^{-n}$, $|x| < 1$

Changing x to $-x$ in (1.10), we get

$$= (1-x)^{-n} = 1 - n(-x) + \frac{n(n+1)}{r!}(-x)^2 + ...$$

$$+ (-1)^r \frac{n(n+1)(n+2)....(n+r-1)}{r!}.(-1)^r x^r + ... \text{to } \infty$$

$$(1+x)^{-2} = 1-2x+3x^2-4x^3+5x^4 - \ldots.$$

$$(1+x)^{-3} = 1-3x+6x^2-10x^3+15x^4 - \ldots.$$

Similarly, putting n =1,2 and 3 successively in (1.11), we have,

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

$$(1-x)^{-2} = 1+2x+3x^2+4x^3+5x^4 + \ldots.$$

$$(1-x)^{-3} = 1+3x+6x^2+10x^3+15x^4 + \ldots.$$

Remark: In the valid expansion of $(1+x)^{-n}$, the terms are alternatively positive and negative while all the terms are positive in $(1-x)^{-n}$. But in the expansion of $(1+x)^n$, when 'n' is a positive fractional value, no such general rule can be stated. For, in the expansion of $(1+x)^{3/2}$, $|x|<1$, we have,

$$(1+x)^{\frac{3}{2}} = 1+\frac{3}{2}x+\frac{\frac{3}{2}\left(\frac{3}{2}-1\right)}{2!}x^2+\frac{\frac{3}{2}\left(\frac{3}{2}-1\right)\left(\frac{3}{2}-2\right)}{3!}x^3+\ldots.$$

$$=1+\frac{3}{2}x+\frac{3}{8}x^2-\frac{1}{16}x^3+\ldots..$$

while in the expansion of $(1+x)^{5/4}$, $|x|<1$, we have

$$(1+x)^{5/8} = 1+\frac{5}{8}x-\frac{15}{128}x^2+\ldots.$$

1.6.9: Sum of a series:

In some specific cases, Binomial Theorem can be used to determine the value of an infinite series provided, the series conforms to the binomial expansion. We explain the method with the help of the following example.

We know that, for any index 'n',

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \ldots \qquad (2)$$

If the series in question conforms to binomial expansion, then each term of (1) must be equal to the corresponding term on the RHS of (2).

Equating the terms of (1) with the corresponding term of (2) we have,

$$nx = \frac{2}{3}\cdot\frac{1}{2} = \frac{1}{3} \qquad (3)$$

$$\frac{n(n-1)}{2!}x^2 = \frac{2.5}{3.6}\cdot\frac{1}{2^2} = \frac{5}{36} \qquad (4)$$

$$\text{or } (nx)^2 - (nx).x = \frac{5}{18}$$

$$\text{or } \left(\frac{1}{3}\right)^2 - \left(\frac{1}{3}\right)x = \frac{5}{18} \qquad \text{[Using (3)]}$$

$$\text{or } \frac{x}{3} = \frac{1}{9} - \frac{5}{18} = -\frac{1}{6}$$

$$\text{or } x = -\frac{1}{2}$$

Putting this value of x in (3) we get,

$$-\frac{1}{2}n = \frac{1}{3}$$

$$\text{or } n = -\frac{2}{3}$$

$$1-\frac{2}{3}\left(-\frac{1}{2}\right)+\frac{\left(-\frac{2}{3}\right)\left(-\frac{2}{3}-1\right)}{2!}\left(-\frac{1}{2}\right)^2+\frac{\left(-\frac{2}{3}\right)\left(-\frac{2}{3}-1\right)\left(-\frac{2}{3}-2\right)}{3!}\left(-\frac{1}{2}\right)^3+\ldots\ldots$$

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

$$=\sqrt[3]{4} \quad \text{as before}$$

Example: 1.46

$$\text{If } y=\frac{1}{3}+\frac{1.3}{3.6}+\frac{1.3.5}{3.6.9}+\frac{1.3.5.7}{3.6.9.12}+\ldots\ldots$$

Prove that $y^2+2y-2=0$

Solution:

The given series may be written as

$$y=\left[1+\frac{1}{3}+\frac{1.3}{3.6}+\frac{1.3.5}{3.6.9}+\ldots\ldots\right]-1$$

Comparing the series in the brackets with

$$1+nx+\frac{n(n-1)}{1.2}x^2+\ \ldots\ldots, \text{ we have}$$

$$nx=\frac{1}{3} \text{ and } \frac{n(n-1)}{1.2}x^2=\frac{1}{3}.\frac{3}{6}$$

Solving for n and x, we get $n=-\frac{1}{2}$ and $x=-\frac{2}{3}$

$\Rightarrow \quad y^2 + 2y - 2 = 0$

Example: 1.47

Using Binomial Theorem, find the value of $\sqrt[3]{126}$ to four decimal places.

Solution:

$$\sqrt[3]{126} = (126)^{\frac{1}{3}} = (125+1)^{\frac{1}{3}} = (125)^{\frac{1}{3}}\left[1+\frac{1}{125}\right]^{\frac{1}{3}}$$

$$= 5[1+0.008]^{\frac{1}{3}} = 5\left[1+\frac{1}{3}\times(0.008)+\frac{\frac{1}{3}\left(-\frac{2}{3}\right)}{2.1}\times(0.008)^2\right] \text{ [neglecting the other terms]}$$

So, $\sqrt[3]{126} = 5\,[1 + 0.002666 - 0.\ 00000711] = 5\,[1.002659]$

$= 5.013295 \ = 5.0133$ (correct to four places of decimal)

Example: 1.48

Find the coefficient of x^{10} in the expansion of $\frac{1+2x}{(1-2x)^2}$, $|x| < \frac{1}{2}$

Solution:

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

$$= (1+2x)\,[\,1 + 2(2x) + 3(2x)^2 + \ldots.+10(2x)^9 + 11(2x)^{10} + \ldots.],\ [\because\ |x| < \frac{1}{2}]$$

$$= (1+2x)\,[\,1 + 4x + 12x^2 + \ldots.+10.\ 2^9.x^9 + 11\ .\ 2^{10}.x^{10} + \ldots.]$$

that the Binomial expansion is valid.

Solution:

The given expression $\frac{1}{(1+x)^2\sqrt{1+4x}}=(1+x)^{-2}(1+4x)^{-\frac{1}{2}}$

$$=\left[1+\frac{(-2)x}{1!}+\frac{(-2)(-2-1)}{2!}x^2+\;....\right]\times\left[1+\frac{\left(-\frac{1}{2}\right)4x}{1!}+\frac{\left(-\frac{1}{2}\right)\left(-\frac{1}{2}-1\right)}{2!}(4x)^2+.....\right]$$

$=(1-2x+3x^2+..........)\,(1-2x+6x^2+..........)$ (ignoring higher powers of x)

$=1-4x+13x^2$

Example: 1.50

Using Binomial theorem, expand $\frac{1}{\sqrt{5+4x}}$ in ascending powers of x.

Solution;

$$\frac{1}{\sqrt{5+4x}}=(5+4x)^{-\frac{1}{2}}=\left[5\left(1+\frac{4x}{5}\right)\right]^{-\frac{1}{2}}=5^{-\frac{1}{2}}\left(1+\frac{4x}{5}\right)^{-\frac{1}{2}}$$

$$=\frac{1}{\sqrt{5}}\left[\left(1+\frac{4x}{5}\right)\right]^{-\frac{1}{2}}=\frac{1}{\sqrt{5}}\left[1+\left(-\frac{1}{2}\right)\left(\frac{4x}{5}\right)+\frac{\left(-\frac{1}{2}\right)\left(-\frac{1}{2}-1\right)}{2!}\left(\frac{4x}{5}\right)^2\right.$$

$$=\frac{}{\sqrt{5}}\left[1-\frac{}{5}+\left(\frac{}{2}\times\frac{}{2}\times\frac{}{2}\times\frac{}{25}\right)-\left(\frac{}{6}\times\frac{}{2}\times\frac{}{2}\times\frac{}{2}\times\frac{}{125}\right)+\dots\right]$$

$$=\frac{1}{\sqrt{5}}\left(1-\frac{2x}{5}+\frac{6x^2}{25}-\frac{4x^3}{25}+\dots\right)$$

Example: 1.51

Prove that $(1+x+x^2+x^3\dots)(1-x+x^2-x^3+\dots)=1+x^2+x^4+x^6+\dots$

Solution:

$$(1+x+x^2+x^3\dots)(1-x+x^2-x^3+\dots)$$

$$=(1-x)^{-1}(1+x)^{-1}=[(1-x)((1+x)]^{-1}=(1-x^2)^{-1}$$

$$=1+x^2+(x^2)^2+(x^2)^3+\dots$$

$$=1+x^2+x^4+x^6+\dots$$

***

2. (a) Find the value of $(630)^{1/4}$ correct to five places of decimals.

(b) Find the 5th root of 244 correct to three places of decimals.

3. (a) Find the coefficient of x^5 in the expansion of $(1-2x)^{-5/2}$, $|x|<\frac{1}{2}$.

(b) Expand $\left(1+4x+x^2\right)^{1/2}$ up to term involving x^3. For what value of 'x' is the expansion valid?

(c) Write the first three terms in the expansion of $\frac{2+x}{(3-2x)^2}$ in ascending powers of x.

Also state the condition under which the result is valid.

4. Show that the general term in the expansion of $(1-2x)^{-\frac{1}{2}}$ is $\frac{(2r)!}{2^r(r!)^2}x^r$.

5. Find the coefficient of x^{10} in $\frac{1+3x^2}{\left(1-x^2\right)^3}$ mentioning the conditions under which the result holds.

6. Assuming 'x' to be small, so that x^2 and higher powers of x are neglected, find the value of $(1-2x)^{\frac{2}{3}}(4+5x)^{\frac{3}{2}}/\sqrt{1-x}$

7. Show that

(i) $x^n=1+n\left(1-\frac{1}{x}\right)+\frac{n(n+1)}{2!}\left(1-\frac{1}{x}\right)^2+\text{....to }\infty$

(ii) $(1+x)^n=2^n\left[1-n\left(\frac{1-x}{1+x}\right)+\frac{n(n+1)}{2!}\left(\frac{1-x}{1+x}\right)^2+\text{....to }\infty\right]$

8. (a) Find the 5th term in the expansion of $\left(1-x^3\right)^{\frac{}{2}}$.

(b) Find the coefficient of x^2 in the expansion of $\dfrac{1+2x}{(1-3x)^3}$

(c) Show that the coefficient of y^n in the expansion of $\dfrac{(1+y)^2}{(1-y)^2}$ is 4n.

(d) Find the coefficient of x^6 in $(1+x+x^2)^{-3}$.

$$\left[\text{Hint: } 1+x+x^2=\frac{1-x^3}{1-x}\right]$$

9. Show that the coefficient of x^r in the expansion of $(1-4x)^{\frac{1}{2}}$ is $\dfrac{(2r)!}{(r!)^2}$

10. When 'x' is so small that its square and higher powers may be neglected, and if

$$\frac{(1-3x)^{\frac{1}{2}}+(1-x)^{\frac{5}{3}}}{\sqrt{4-x}}$$ is approximately equal to a+bx, find a and b.

11. If n > 4 and cubes and higher powers of $\dfrac{1}{n}$ are neglected, show that

$$\sqrt{n^2+16}-\sqrt{n^2+9}=\frac{7}{2n}$$

12. If the binomial expansion of $(a+bx)^{-2}$ is $\dfrac{1}{4}-3x+\dots$, find the values of a and b.

13. Use the binomial theorem to evaluate

(v) $\frac{1}{\sqrt[3]{8.16}}$ correct to four decimal places.

14. With the help of the Binomial Theorem, show that

$$1+\frac{1}{4}+\frac{1}{4}.\frac{3}{8}+\frac{1.3.5}{4.8.12}+\frac{1.3.5.7}{4.8.12.16}+....=\sqrt{2}$$

15. Identify the following series as binomial expansion and hence find its sum

$$1+\frac{7}{18}+\frac{7.9}{18.36}+\frac{7.9.11}{18.36.54}+....\text{to } \infty$$

16. Sum the series

$$2+\frac{5}{2!3}+\frac{5.7}{3!3^2}+\frac{5.7.9}{4!3^3}+....$$

[Hint : Write $2 = 1+1 = 1+\frac{\frac{3}{2}}{1}.\frac{2}{3}$

So the series becomes $1+\frac{\frac{3}{2}}{1}.\frac{2}{3}+\frac{\frac{3}{2}.\frac{5}{2}}{2!}\frac{2^2}{3^2}+\frac{\frac{3}{2}.\frac{5}{2}.\frac{7}{2}}{3!}.\frac{2^3}{3^3}+....$]

17. If a and b are values of the second and third terms respectively in the expansion of $(1+x)^n$, prove that

$$n=\frac{a^2}{a^2-2b} \text{ and } x=\frac{a^2-2b}{a}$$

18. When x is so small that its square and higher powers can be neglected, show that

(ii) The coefficient of x in the expansion of $(1+3x)$ $(1-2x)$ is ________.

(iii) If $y = 2x + 3x^2 + 4x^3 + \ldots$, then $x =$ ________.

(iv) The value of $\sqrt[3]{998}$, correct to four places of decimals is ________.

20. **Choose the correct alternative :**

(i) $\dfrac{1}{\sqrt{4-3x^2}}$ can be expanded in ascending powers of x by using Binomial Theorem if

(a) $x < \dfrac{2}{\sqrt{3}}$, (b) $x > \dfrac{2}{\sqrt{3}}$, (c) $-\dfrac{2}{\sqrt{3}} < x < \dfrac{2}{\sqrt{3}}$, (d) $|x| > \dfrac{2}{\sqrt{3}}$

(ii) The coefficient of x^2 in the expansion of $(1 + 4x + x^2)^{1/2}$ is

(a) - 3, (b) - 2, (c) 2, (d) none of these

(iii) $\dfrac{1}{\sqrt[3]{128}}$ is equal to

(a) 0.1984 (b) 0.1993 (c) 0.2001 (d) none of these

(iv) The sum of the series.

$$1 + \frac{3}{4} + \frac{3}{4}.\frac{5}{8} + \frac{3}{4}.\frac{5}{8}.\frac{7}{12} + \ldots \text{ is}$$

(a) $\sqrt{2}$ (b) $2\sqrt{2}$ (c) 2^n (d) none of these

***

is usually denoted by 'e' and is of fundamental importance in Mathematics. Although the series denoted by 'e' is infinite, its value is finite and lies between 2 and 3.

(i) Prove that $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e$

Proof : Expanding $\left(1+\frac{1}{n}\right)^n$ by the Binomial Theorem when n >1, we have

$$\left(1+\frac{1}{n}\right)^n = 1+n\,.\frac{1}{n}+\frac{n(n-1)}{2!}\left(\frac{1}{n}\right)^2+\frac{n(n-1)(n-2)}{3!}\left(\frac{1}{n}\right)^3+....$$

$$=1+\frac{1}{1!}+\frac{1}{2!}\,.\left(1-\frac{1}{n}\right)+\frac{1}{3!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)+....$$

Now, as $n\to\infty$, $\frac{1}{n},\frac{2}{n},\frac{3}{n}$ etc. all tend to zero.

$$\therefore\ \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = 1+1+\frac{1}{2!}+\frac{1}{3!}+....= e$$

Hence $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e$

(ii) Show that, the value of e lies between 2 and 3.

Proof : From definition, we know that

$$e = 1+1+\frac{1}{2!}+\frac{1}{3!}+....$$

$$= 2+\frac{1}{2!}+\frac{1}{3!}+....$$

which is greater than 2.

Hence, $e = 1+1+\frac{1}{2!}+\frac{1}{3!}+\ldots. < 1+1+\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^4}+\ldots.$

$$= 1+\frac{1}{1-\frac{1}{2}} = 1+2 = 3$$

So, $2 < e < 3$

(iii) Show that the approximate value of 'e' up to four places of decimals is 2.7183.

Proof : By definition, we have,

$$e = 1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\frac{1}{4!}+ \ldots\ldots \text{to } \infty$$

$$= 2+\frac{1}{2!}+\frac{1}{3!}+\frac{1}{4!}+ \ldots\ldots \text{to } \infty$$

Now, $2 = 2.000000,\ \frac{1}{2!} = \frac{1}{2} = 0.500000,\ \frac{1}{3!} = \frac{1}{3.2!} = \frac{1}{3}(0.500000) = 0.166667$

$$\frac{1}{4!} = \frac{1}{4.3!} = \frac{1}{4}(0.166667) = 0.041667,\ \frac{1}{5!} = \frac{1}{5.4!} = \frac{1}{5}(0.041667) = 0.008333$$

$$\frac{1}{6!} = \frac{1}{6.5!} = \frac{1}{6}(0.008333) = 0.001389,\ \frac{1}{7!} = \frac{1}{7.6!} = \frac{1}{7}(0.001389) = 0.000198$$

$$\frac{1}{8!} = \frac{1}{8.7!} = \frac{1}{8}(0.000198) = 0.000025,\ \frac{1}{9!} = \frac{1}{9.8!} = \frac{1}{9}(0.000025) = 0.000003$$

$\therefore$ $e = 2.000000 + 0.500000 + 0.166667 + 0.041667 + 0.008333 + 0.001389 +$
$0.000198 + 0.000025 + 0.000003 = 2.7183$

Multiplying both sides by n!, we get

$$(n-1)!m = n! + n! + \frac{n!}{2!} + \frac{n!}{3!} + \ldots + 1 + \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \ldots$$

Since 'n' is a positive integer, (n -1)! is also a positive integer and therefore (n - 1)! m is also a positive integer.

So, the RHS must be a positive integer.

But, the RHS = $n! + n! + \frac{n!}{2!} + \frac{n!}{3!} + \ldots + 1 + \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \ldots$

$$= \text{a positive integer } + \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \ldots$$

$$\therefore \quad (n-1)!m = \text{a positive integer} + \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \frac{1}{(n+1)(n+2)(n+3)} + \ldots$$

$$= \text{a positive integer} + s \qquad (1.12)$$

where, $s = \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \frac{1}{(n+1)(n+2)(n+3)} + \ldots$

It is evident that $s > \frac{1}{n+1}$ (1.13)

Now the first term of 's' is $\frac{1}{n+1}$

The 2nd term = $\frac{1}{(n+1)(n+2)} < \frac{1}{(n+1)(n+1)} = \frac{1}{(n+1)^2}$

The 3rd term = $\frac{1}{(n+1)(n+2)(n+3)} < \frac{1}{(n+1)(n+1)(n+1)} = \frac{1}{(n+1)^3}$ etc.

$$= \frac{1}{n+1}\left[1 - \frac{1}{n+1}\right]^{-1}$$

$$= \frac{1}{n+1}\left(\frac{n+1}{n}\right)$$

$$= \frac{1}{n} \qquad (1.14)$$

From (1.13) and (1.14), we get, $\frac{1}{n+1} < s < \frac{1}{n}$

i.e. 's' is a positive fraction.

But this is impossible because the RHS of (1.12) must be a positive integer.

Thus the assumption that 'e' is commensurable does not hold.

Hence, 'e' is incommensurable -

Note : This method is also known as the method of reduction and absurdum.

1.7.1. Exponential Theorem :

If x is any real quantity and $a > 0$, then

(i) $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \text{to } \infty$

(ii) $a^x = 1 + x \log a + \frac{(x\log a)^2}{2!} + \frac{(x\log a)^3}{3!} + \dots \text{to } \infty$, where the base of log is e.

Proof : (i) By Binomial Theorem we know that,

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \dots \text{to } \infty$$

Expanding by Binomial Theorem, when 'n' is greater than 1, we have,

$$\left(1+\frac{1}{n}\right)^{nx}=1+nx.\frac{1}{n}+\frac{nx(nx-1)}{2!}.\frac{1}{n^2}+\frac{nx(nx-1)(nx-2)}{3!}.\frac{1}{n^3}+....$$

$$=1+x+\frac{nx}{n}\frac{(nx-1)}{n}.\frac{1}{2!}+\frac{nx}{n}\frac{(nx-1)}{n}\frac{(nx-2)}{n}.\frac{1}{3!}+....$$

$$=1+x+x\left(x-\frac{1}{n}\right)\frac{1}{2!}+x\left(x-\frac{1}{n}\right)\left(x-\frac{2}{n}\right)\frac{1}{3!}+....$$

Now taking limit $n\to\infty$ on both sides and using the fact that, $\frac{1}{n},\frac{2}{n},\frac{3}{n},....\to 0$ as $n\to\infty$, we have,

$$\underset{n\to\infty}{\text{Lt}}\left[\left(1+\frac{1}{n}\right)^{n}\right]^{x}=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+....$$

$$\text{i.e.}\left[\underset{n\to\infty}{\text{Lt}}\left(1+\frac{1}{n}\right)^{n}\right]^{x}=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+....$$

Hence, $e^{x}=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+....\text{to }\infty$

(ii) Let $y=a^x$, then $\log_e y=x\ \log_e a$

Hence by the definition of logarithm, we have

$y=e^{x\log_e a}$ (for if $e^m=n$, then $m=\log_e n$)

$\therefore\ a^x=e^{x\log_e a}$

But $e^{x\log_e a} = a^x$

So, $a^x = 1 + x\log_e a + \frac{(x\log_e a)^2}{2!} + \frac{(x\log_e a)^3}{3!} + \ldots.$

Some particular Cases :

The exponential series is $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots.$

1. When $x = 0$, e^x becomes $e^0 = 1+0+0+\ldots.. = 1$

2. When $x = 1$, e^x becomes $e^1 = 1 + \frac{1}{1!} + \frac{1}{2!} + \ldots. = e$

3. When $x = -1$, e^x becomes $e^{-1} = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots.$

Thus, $\left(1 + \frac{1}{1!} + \frac{1}{2!} + \ldots\right)^{-1} = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots.$

4. Changing x to $-x$ in e^x we get,

$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \ldots.$ to ∞

The terms have alternately positive and negative signs in this series.

Students are advised to remember the following useful expansions :

(i) $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ \ldots. \ + \frac{x^n}{n!} + \ \ldots.$

(ii) $e^{-x} = 1 - \frac{x}{1!} + \frac{x^2}{2!} - \frac{x^3}{3!} + \ \ldots. \ + (-1)^n \frac{x^n}{n!} + \ \ldots.$

(vi) $e^x - e^{-x} = 2\left(\frac{x}{1!} + \frac{x^3}{3!} + \frac{x^5}{5!} + \dots\right)$

Example : 1.52 Find the value of e^2 rounded off to one place of decimals.

Solution : Putting $x = 2$ in the expansion of e^x, we have,

$$e^2 = 1 + \frac{2}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!} + \frac{2^5}{5!} + \frac{2^6}{6!} + \dots\dots$$

$$= 1 + 2 + 2 + \frac{4}{3} + \frac{2}{3} + \frac{4}{15} + \frac{4}{45} + \dots\dots$$

> 7.355; the sum of the first seven terms of e^2

Also, $$e^2 < \left(1 + \frac{2}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!}\right) + \frac{2^5}{5!}\left(1 + \frac{2}{6} + \frac{2^2}{6^2} + \frac{2^3}{6^3} + \dots\right)$$

(taking 6! in places of 7!, 8!, etc. in rest of the terms)

$$= \left(1 + 2 + 2 + \frac{4}{3} + \frac{2}{3}\right) + \frac{2^5}{5!}\left(1 + \frac{1}{3} + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^3 + \dots\right)$$

$$= 7 + \frac{2^5}{5!}\left[1 - \frac{1}{3}\right]^{-1}$$

$$= 7 + \frac{4}{15} \cdot \frac{3}{2} = 7 + 0.4$$

i.e. $e^2 < 7.4$.

Thus, $7.355 < e^2 < 7.4$

Hence, the coefficient of x^n in e^{a+bx} is $e^a.\frac{b^n}{n!}$

Example : 1.54 Write the series for $\left(e^{3x}-e^{-3x}\right)/\ 2$

Solution : We know that $e^x=1+\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+....$

Changing x to 3x we get,

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

Again changing x to $-$ 3x, we get

$$e^{-3x}=1+\frac{(-3x)}{1!}+\frac{(-3x)^2}{2!}+\frac{(-3x)^3}{3!}+....$$

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

On subtraction, we get

$$e^{3x}-e^{-3x}=2\left\{\frac{(3x)}{1!}+\frac{(3x)^3}{3!}+\frac{(3x)^5}{5!}+....\right\}$$

$$\therefore\quad \frac{e^{3x}-e^{-3x}}{2}=\frac{3x}{1!}+\frac{(3x)^3}{3!}+\frac{(3x)^5}{5!}+...$$

Example : 1.55 Find the sum of $1+\frac{1}{2!}+\frac{1}{4!}+\frac{1}{6!}+....$

Solution : We know that $e^x=1+\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+....$

$$\therefore \quad 1+\frac{1}{2!}+\frac{1}{4!}+\frac{1}{6!}+\ldots .=\frac{1}{2}\left(e+e^{-1}\right) \ =\frac{1}{2}\left(e+\frac{1}{e}\right)$$

Example : 1.56 Sum the series $\sum\limits_{n=1}^{\infty} \frac{2n}{n!}$

Solution : Here the general term $t_n=\frac{2n}{n!}=\frac{2n}{n(n-1)!}=\frac{2}{(n-1)!}$

$$\therefore \quad \sum_{n=1}^{\infty} \frac{2n}{n!}=\sum_{n=1}^{\infty} \frac{2}{(n-1)!}=2 \sum_{n=1}^{\infty} \frac{1}{(n-1)!}$$

$$=2\left[\frac{1}{(1-1)!}+\frac{1}{(2-1)!}+\frac{1}{(3-1)!}+\frac{1}{(4-1)!}+\ldots .\right]$$

$$=2\left[1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\ldots .\right]=2e$$

Example : 1.57 Sum the Series $1+\frac{2^2}{2!}+\frac{3^2}{3!}+\frac{4^2}{4!}+\ldots .$

Solution : Here $t_n=\frac{n^2}{n!}=\frac{n}{(n-1)!}$

Putting n = 1, 2, 3, 4,....., successively, we have,

$$t_1=\frac{1}{(1-1)!}, \quad t_2=\frac{2}{(2-1)!}, \quad t_3=\frac{3}{(3-1)!}, \quad t_4=\frac{4}{(4-1)!}$$

Adding, we get,

$$s=\sum_{n=1}^{\infty} t_n=\sum_{n=1}^{\infty} \frac{n}{(n-1)!}$$

$$=\left[1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+....\right]+\left[\frac{1}{1!}+\frac{2}{2!}+\frac{3}{3!}+....\right]$$

$$=e+\left(1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+....\right)=e+e=2e$$

Alt. $t_n=\frac{n^2}{n!}=\frac{n(n-1)+n}{n!}=\frac{1}{(n-2)!}+\frac{1}{(n-1)!}$

$$s=\sum_{n=1}^{\infty}t_n=\sum_{n=1}^{\infty}\frac{1}{(n-2)!}+\sum_{n=1}^{\infty}\frac{1}{(n-1)!}$$

$$=\sum_{n=2}^{\infty}\frac{1}{(n-2)!}+\sum_{n=1}^{\infty}\frac{1}{(n-1)!}$$

$$=e+e=2e$$

Example : 1.58 Show that $1+\frac{1}{2!}+\frac{1.3}{4!}+\frac{1.3.5}{6!}+....=\sqrt{e}$

Solution : Let t_n be the n th term of the given series.

Then $t_n=\frac{1.3.5.....(2n-3)}{(2n-2)!}$

$$=\frac{1.2.3.4.5.6.....(2n-3)(2n-2)}{(2n-2)!\,2.4.6....(2n-2)}\text{ for } n\geq 2$$

$$\therefore\quad t_n=\frac{(2n-2)!}{(2n-2)!.2^{n-1}.(n-1)!}=\left(\frac{1}{2}\right)^{n-1}\frac{1}{(n-1)!}\qquad .....(1)$$

Clearly t_1, the 1st term of the series = 1

$$1+\frac{\frac{1}{2}}{1!}+\frac{\left(\frac{1}{2}\right)^2}{2!}+\frac{\left(\frac{1}{2}\right)^3}{3!}+\ldots.$$

$$=e^{\frac{1}{2}}=\sqrt{e}$$

***

EXERCISES - 1.5

1. (a) Show that the coefficient of x^{10} in the expansion of e^{2x} is $\frac{4}{14175}$.

   (b) Find the coefficient of x^{12} in the expansion of e^{2x}.

2. Write the series for.

   (i) e^{4x} (ii) $e^{2x}+e^{-2x}$ (iii) $\frac{e^{x}-e^{-x}}{2}$ (iv) $\frac{e+e^{-1}}{2}$ (v) $\frac{e-e^{-1}}{2}$

3. (a) Prove that $\frac{2}{1!}+\frac{4}{3!}+\frac{6}{5!}+\ldots.=e$

   (b) Find the coefficient of x^{n} in the expansion of $\frac{a-bx}{e^{x}}$.

   (c) Prove that the coefficient of x^{n} in

   $$\frac{1+x}{1!}+\frac{(1+x)^{2}}{2!}+\frac{(1+x)^{3}}{3!}+\ldots.\text{is }\frac{e}{n!}.$$

4. (a) Show that $\frac{e^{ix}+e^{-ix}}{2}=1-\frac{x^{2}}{2!}+\frac{x^{4}}{4!}+\ldots.$, where $\sqrt{-1}=i$

5. Show that

(a) $\frac{e^{ix}-e^{-ix}}{2}=x-\frac{x^3}{3!}+\frac{x^5}{5!}+....$

(b) $\left(1+\frac{1}{2!}+\frac{1}{4!}+....\right)^2-\left(1+\frac{1}{3!}+\frac{1}{5!}+....\right)^2=1$

6. (a) Find the coefficient of a^n in the series

$$\frac{1+a}{1!}+\frac{(1+a)^2}{2!}+\frac{(1+a)^3}{3!}+....$$

(b) Find the sum of the infinite series

$$1+\frac{2^3}{2!}+\frac{3^3}{3!}+\frac{4^3}{4!}+....$$

(c) Show that, $1+\frac{1+3}{2!}+\frac{1+3+5}{3!}+\frac{1+3+5+7}{4!}+....=2e$

7. Sum the series from n = 1 to ∞, whose n th term is

(i) $\frac{1}{(n+1)!}$, (ii) $\frac{1}{(n+2)!}$, (iii) $\frac{1}{(2n-1)!}$, (iv) $\frac{1}{(2n+1)!}$, (v) $\frac{n^2}{(n+1)!}$

(vi) $\frac{2n}{n!}$, (vii) $\frac{(2n)!}{n!}$, (viii) $\frac{2^n}{n!}$, (ix) $\frac{C(n,\,2)}{(n+1)!}$

(x) $\frac{C(n,o)+C(n,1)+....+C(n,n)}{P(n,n)}$

(ii) $\frac{2}{3!}+\frac{4}{5!}+\frac{6}{7!}+\frac{8}{9!}+....$

1.8 LOGARITHMIC SERIES

Prove that, $\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+.....$, when $|x|<1$.

Proof : By the previous section 1.7.1 (ii) we know that

$$a^x=1+x\log_e a+\frac{\left(x\log_e a\right)^2}{2!}+\frac{\left(x\log_e a\right)^3}{3!}+....=e^{x\log_e a}$$

So, $a^y=e^{y\log_e a}$

$$=1+y\log_e a+\frac{y^2\left(\log_e a\right)^2}{2!}+\frac{y^3\left(\log_e a\right)^3}{3!}+....$$

Writing a = 1+x in this series, we have,

$$(1+x)^y=1+y\log_e(1+x)+\frac{y^2}{2!}[\log_e(1+x)]^2+\frac{y^3}{3!}[\log_e(1+x)]^3+...... \quad (1.15)$$

But by the binomial theorem, when x<1, we have

$$(1+x)^y=1+yx+\frac{y(y-1)}{2!}x^2+\frac{y(y-1)(y-2)}{3!}x^3+.... \quad (1.16)$$

Since the LHS of (1.15) and (1.16) are equal, their RHS must be equal

i.e. $1+y\log_e(1+x)+\frac{y^2}{2!}[\log_e(1+x)]^2+....=1+yx+\frac{y(y-1)}{2!}x^2+\frac{y(y-1)(y-2)}{3!}x^3+....(1.17)$

Equating the coefficient of y on both sides of (1.17), we have, This is called logarithmic series

Note : Expansion of $\log_e(1+x)$, is also valid if $x = 1$. The proof is not given here as it is beyond the scope of the present text. It is taken for granted that the expansion of $\log_e(1+x)$ as a power series of x is valid for $-1 < x \le 1$

Cor.1. Changing x to - x in the logarithmic series (1.18), we have

$$\log_e(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \ldots.$$

$$= -\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \ldots.\right) \qquad (1.19)$$

Deductions :

(i) Subtracting (1.19) from (1.18), we have

$$\log_e(1+x) - \log_e(1-x) = 2\left[x + \frac{x^3}{3} + \frac{x^5}{5} + \ldots\right]$$

$$\text{or,} \quad \frac{1}{2}\log_e\frac{1+x}{1-x} = x + \frac{x^3}{3} + \frac{x^5}{5} + \ldots.., (\,|x|<1\,)$$

(ii) Putting $x = 1$ in logarithmic series,

$$\log_e 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \ldots.$$

Remarks :

(a) The series expansion for $\log_e(1+x)$ may fail to be valid if $|x| > 1$.

(b) There are three major differences between the exponential series and the logarithmic series. Those are stated below :

(i) In the series $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots.$, all terms carry positive signs, while in the series

Example : 1.59 Prove that $\frac{1}{2}-\frac{1}{2}\cdot\frac{1}{2^2}+\frac{1}{3}\cdot\frac{1}{2^3}-\frac{1}{4}\cdot\frac{1}{2^4}+....=\log\frac{3}{2}$

Solution : LHS = $\frac{1}{2}-\frac{1}{2}\cdot\frac{1}{2^2}+\frac{1}{3}\cdot\frac{1}{2^3}-\frac{1}{4}\cdot\frac{1}{2^4}+....$

Putting $x=\frac{1}{2}$,

$$\text{L.H.S} = x-\frac{1}{2}x^2+\frac{1}{3}x^3-\frac{1}{4}x^4+....$$

$$=\log(1+x)=\log\left(1+\frac{1}{2}\right)=\log\frac{3}{2}=\text{R.H.S.}$$

Example : 1.60

Using the logarithmic series, prove that the value of log 2 lies between 0.61 and 0.76

Solution :

$$\log 2=\log(1+1)=1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\frac{1}{6}+....$$

$$=\left(1-\frac{1}{2}\right)+\left(\frac{1}{3}-\frac{1}{4}\right)+\left(\frac{1}{5}-\frac{1}{6}\right)+....=\frac{1}{2}+\frac{1}{12}+\frac{1}{30}+....$$

= (0.5+0.083+0.033) + positive terms = 0.616 + (sum of positive terms)

$\therefore$ $\log 2 > 0.616$

Again, $\log 2=1-\left(\frac{1}{2}-\frac{1}{3}\right)-\left(\frac{1}{4}-\frac{1}{5}\right)-\left(\frac{1}{6}-\frac{1}{7}\right)-\left(\frac{1}{8}-\frac{1}{9}\right)....$

$$=1-\frac{1}{6}-\frac{1}{20}-\frac{1}{42}-\frac{1}{72}-....$$

Hence the result.

Example : 1.61

If $y = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots.,$ and if $|x| < 1$, prove that

$$x = y + \frac{y^2}{2!} + \frac{y^3}{3!} + \frac{y^4}{4!} + \ldots.$$

Solution :

$$y = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots. = \log_e(1+x)$$

By definition of logarithm,

$$y = \log_e(1+x) \Rightarrow 1+x = e^y$$

$$\therefore \quad x = e^y - 1$$

$$= \left(1 + y + \frac{y^2}{2!} + \frac{y^3}{3!} + \ldots.\right) - 1$$

$$= y + \frac{y^2}{2!} + \frac{y^3}{3!} + \ldots.$$

Example : 1.62

Prove that $\log\sqrt{2} = \frac{1}{2}\left(\frac{1}{2} + \frac{1}{3}\right) - \frac{1}{4}\left(\frac{1}{2^2} + \frac{1}{3^2}\right) + \frac{1}{6}\left(\frac{1}{2^3} + \frac{1}{3^3}\right)$

Solution :

$$\text{R.H.S} = \frac{1}{2}\left(\frac{1}{2} + \frac{1}{3}\right) - \frac{1}{4}\left(\frac{1}{2^2} + \frac{1}{3^2}\right) + \frac{1}{6}\left(\frac{1}{2^3} + \frac{1}{3^3}\right)$$

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

$$=\frac{1}{2}\left(\log\frac{3}{2}+\log\frac{4}{3}\right)=\frac{1}{2}\ \log\ \left(\frac{3}{2}\times\frac{4}{3}\right)$$

$$=\frac{1}{2}\log 2\quad=\log 2^{\frac{1}{2}}\quad=\log\sqrt{2}\quad=\text{L.H.S.}$$

Example : 1.63

Find the sum of $\frac{5}{1.2.3}+\frac{7}{3.4.5}+\frac{9}{5.6.7}+....$

Solution :

The nth term, $t_n=\frac{2n+3}{(2n-1)2n(2n+1)}$ .... (1)

The RHS of (1) can be reduced in to partial fractions as follows :

Let $\frac{2n+3}{(2n-1)2n(2n+1)}=\frac{A}{2n-1}+\frac{B}{2n}+\frac{C}{2n+1}$ .... (2)

where A, B and C are constants. Multiplying both sides by (2n – 1) 2n (2n+1), we get

$A(2n)\ (2n+1)+B\ (2n-1)\ (2n+1)+C\ (2n\ -1)\ 2n=2n+3$

Comparing the coefficients of n^2, 'n' and the constant terms successively on both the sides of (2) we get,

$4A+4B+4C=0,\qquad 2A-2C=2,\quad -B=3$

Solving, we get, $A=2,\ B=-3$ and $C=1$

Substituting these values in (2) we have,

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

Thus the given series $\frac{5}{1.2.3}+\frac{7}{3.4.5}+\frac{9}{5.6.7}+....$

two terms in each of the brackets together and adding, we have,

$$\frac{5}{1.2.3}+\frac{7}{3.4.5}+\frac{9}{5.6.7}+\ldots.$$

$$=2\left(1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\ldots.\right)+\left(-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}\ldots.\right)$$

$$=2\log_e 2+\log_e 2-1$$

$$=3\log_e 2-1$$

$$=\log_e 8-1$$

EXERCISES 1.6

1. Find the sum of the following infinite series.

(i) $\frac{1}{5}+\frac{1}{2}.\left(\frac{1}{5}\right)^2+\frac{1}{3}\left(\frac{1}{5}\right)^3+\ldots.$

(ii) $1+\frac{1}{3}\left(\frac{1}{2}\right)^2+\frac{1}{5}.\left(\frac{1}{2}\right)^4+\ldots.$

2. Prove that $\left(\frac{a-b}{a}\right)+\frac{1}{2}\left(\frac{a-b}{a}\right)^2+\frac{1}{3}\left(\frac{a-b}{a}\right)^3+\ldots.=\log\frac{a}{b}$

3. Show that $\log 10=3\log 2+\frac{1}{4}-\frac{1}{2}\left(\frac{1}{4}\right)^2+\frac{1}{3}\left(\frac{1}{4}\right)^3-\ldots.$

4. Prove that $\log(n+1)-\log n=2\left[\frac{1}{2n+1}+\frac{1}{3(2n+1)^3}+\frac{1}{5(2n+1)^5}+\ldots.\right]$

5. Prove the following :

(c) $\frac{1}{1.2.3}+\frac{1}{3.4.5}+\frac{1}{5.6.7}+....=\log 2-\frac{1}{2}$

(d) $\log(1+x)^{1+x}(1-x)^{1-x}=2\left[\frac{x^2}{1.2}+\frac{x^4}{3.4}+\frac{x^6}{5.6}+....\right]$

(e) $\log\sqrt{12}=1+\left(\frac{1}{2}+\frac{1}{3}\right)\frac{1}{4}+\left(\frac{1}{4}+\frac{1}{5}\right)\frac{1}{4^2}+\left(\frac{1}{6}+\frac{1}{7}\right)\frac{1}{4^3}+....+\text{ to }\infty$

(f) $\frac{x-1}{x+1}+\frac{1}{2}.\frac{x^2-1}{(x+1)^2}+\frac{1}{3}.\frac{x^3-1}{(x+1)^3}+....=\log x,\quad x>0$

(g) $\frac{2}{1}x+\frac{3}{2}x^2+\frac{4}{3}x^3+\frac{5}{4}x^4+....=\frac{x}{1-x}-\log(1-x),\quad |x|<1$

(h) $\frac{1}{2}x^2+\frac{2}{3}x^3+\frac{3}{4}x^4+\frac{4}{5}x^5+....=\frac{x}{1-x}+\log(1-x),\ |x|<1$

(i) $\log 2+2\left[\frac{1}{5}+\frac{1}{3}.\frac{1}{5^2}+\frac{1}{5}.\frac{1}{5^3}+....\text{to }\infty\right]=\log 3$

(j) $\log_3 e-\log_9 e+\log_{27} e-....=\log_3 2$

6. If $y=-x^3-\frac{x^6}{2}-\frac{x^9}{3}-....$, prove that $x^3=1-e^y$

7. Prove that the series $\frac{1}{n+1}+\frac{1}{2(n+1)^2}+\frac{1}{3(n+1)^3}+....$ has the same sum as the series $\frac{1}{n}-\frac{1}{2n^2}+\frac{1}{2n^3}+....$

8. If $y=2x^2-1$, then prove that, under certain conditions (to be stated),

$$\frac{1}{x^2}+\frac{1}{2x^4}+\frac{1}{3x^6}+....=\frac{2}{y}+\frac{2}{3y^3}+\frac{2}{5y^5}+....$$

Hence show that $\log_e 3 = 1+\frac{1}{3.2^2}+\frac{1}{5.2^4}+....$

12. Show that the coefficient of x^n in the expansion of $\log(1+x+x^2)$ is $-\frac{2}{n}$ or $\frac{1}{n}$ according as 'n' is a multiple of 3 or is not a multiple of 3

13. If α and β be the roots of the equation $ax^2+bx+c=0$, show that,

$$\log_e(a-bx+cx^2)=\log_e a+(\alpha+\beta)x-\left(\frac{\alpha^2+\beta^2}{2}\right)x^2+\left(\frac{\alpha^3+\beta^3}{3}\right)x^3+...$$

14. Prove that if

$$f=\frac{x}{1+x^2}+\frac{1}{3}\left(\frac{x}{1+x^2}\right)^3+\frac{1}{5}\left(\frac{x}{1+x^2}\right)^5+....$$

and $g=x-\frac{2}{3}x^3+\frac{1}{5}x^5+\frac{1}{7}x^7-\frac{2}{9}x^9+.....$, then $f = g$.

15. Prove that $\log\left(1+3x+2x^2\right)=3x-\frac{5}{2}x^2+\frac{9}{3}x^3-\frac{17}{4}x^4+....$

16. If $\log_e\frac{1}{1-x-x^2+x^3}$ is expanded in ascending powers of x, show that the coefficient of x^n is $\frac{3}{n}$ if 'n' is even and $\frac{1}{n}$ if 'n' is odd. $(|n|<1)$

17. Show that the coefficient of x^n in the expansion of

$\log_e(1-5x+6x^2)$ is $-\frac{1}{n}(2^n+3^n)$, if $|x| < \frac{1}{3}$

18. For $|x|<1$, if $A = \frac{x}{1+x^2}+\frac{1}{3}\left(\frac{x}{1+x^2}\right)^3+\frac{1}{5}\left(\frac{x}{1+x^2}\right)^5+....$

$x - \frac{\;}{2}x + \frac{\;}{3}x \ldots$

in ascending powers of y.

20. Prove that

$$\frac{1+3}{1!}(\log 3)+\frac{1+3^2}{2!}(\log 3)^2+\frac{1+3^3}{3!}(\log 3)^3+\ldots = 28$$

21. Expand $\log_e(1+x+x^2+x^3)$ in ascending powers of x and find the coefficients of x^{2n} and x^{2n+1}, $(|x|<1)$

***

ANSWERS

EXERCISES 1.1

1. 20 3. (i) n, (ii) n^2+3n+2, 4. (i) 156, (ii) 1/720, 5. 144. 6. 16!x2!, 7. (i) 120, (ii) 40, (iii) 40, (iv) 80, (v) 20, 8. (i) 7!, (ii) 6!, 9. (i) 120, (ii) 24, (iii) 48, 10. (i) 1728, (ii) 576, 11. (i) 9, (ii) 9, (iii) 5 12. 277200, 13. (i) 7560, (ii) 60, 14. 6, 15. 15, 17. 154, 18. 30, 240, 19. 4320, 20. (i) 840, (ii) 560, 21. 1024, 22. 12, 23. (i) 8, (ii) 24, 24. (i) 4, (ii) 1024, (iii) 4320, (iv) 24, (v) 9, (vi) 31, 25. (i) True (ii) False $4^3.5^3$ (iii) False, 2880, (iv) False, 91, (v) True. 26. (i) b, (ii) d, 1281, (iii) a, (iv) c, (v) b, 27. 2700, 28. 63, 29. 2^n, 30. 34650, 31. 5775, 32. (i) 136, (ii) 2454, 33. (i) ${}^5C_3 \times {}^6C_2$ (ii) ${}^5C_2 \times {}^6C_3$ (iii) ${}^5C_5$ (iv) 461, (v)401, 34. ${}^{10}C_1 \times {}^9C_4 \times 2^4$.

(c) $x - 6x + 15x - 20x + 15x - 6x + 1$

2. (a) $2(32x^6 + 48x^4 + 18x^2 + 1)$

(b) $2x^3 + 30x^2y + 30xy^2 + 2y^3$ (c) $58\sqrt{2}$

3. (i) 96059601 (ii) 1126162419264

4. (a) $10500/x^3$ (b) $-5005x^6 y^{18}$ (c) $-1701x^2$ (d) 29568

5. (a) ${}^{12}C_6 a^6 b^6$ (b) $-340x^{10}, \frac{560}{3} x^{11}$

8. (a) 7920 (b) 2268, 9. ±3, 10. 1.0020, 12. (b) - 16, 144

16. (a) 1 or 14 (b) 15, 17. $n = 7, r = 3$, 19. 5, 20. $x = 2, a = 3, n = 5$

21. - 45, 22. 1, 24. (a) 3, (b) ±3 25. (a) $\frac{7}{8}$ (b) ±2

EXERCISES 1.3

1. (i) 2^{12}, (ii) $2^{15} - 1$, 7. 7, 14, 8. 55, 13, 9. 8, 10. $x = 1, a = 2, n = 7$,

11. (a) 12, (b) $a = \frac{1}{2}, n = 8$ 12. (i) 252, (ii) 2, (iii) 101^{50}, (iv) 3^{143} 13. (i) F

(ii) T (iii) T 14. (i) c, (ii) d, 2268, (iii) b, (iv) b,

EXERCISES 1.4

1. (a) $1 - \frac{3}{2}x + \frac{3}{8}x^2 + \frac{1}{16}x^3 + \dots$

(b) $\frac{1}{16}\left(1 - 6x + \frac{45}{2}x^2 - \frac{135}{2}x^3 + \dots\right)$

(c) $\frac{(r+1)(r+2)}{2!}x^r$

5. 66, $0<x^2<1$, 6. $8+\frac{25}{3}x$

8. (a) $\frac{1155}{128}x^{12}$, (b) 72, (d) 3.

10. a = 1, (b) = -35/24 12. a = 2, (b) = 12

13. (i) 9.997, (ii) 0.991, (iii) 2.0025, (iv) 0.994012, (v) 0.4967.

15. $\left(1-\frac{1}{9}\right)^{-7/2}$, 1.50198, 16. $3\sqrt{3}$

19. (i) 18, (ii) $\frac{55}{72}$ (iii) $\frac{y}{2}-\frac{3y^2}{8}+\frac{5y^3}{16}-....$

(iv) 9.9933.

20. (i) c, (ii) b, (iii) a, (iv) b

EXERCISES 1.5

1. (b) $\frac{4}{467775}$ 2.(i) $1+\frac{4x}{1!}+\frac{(4x)^2}{2!}+\frac{(4x)^3}{3!}+.....$

(ii) $2\left(1+\frac{(2x)^2}{2!}+\frac{(2x)^4}{4!}+....\right)$ (iii) $x+\frac{x^3}{3!}+\frac{x^5}{5!}+....$

(iv) $1+\frac{1}{2!}+\frac{1}{4!}+....$ (v) $\frac{1}{1!}+\frac{1}{3!}+\frac{1}{5!}+....$

3. (b) $\frac{(-1)^n}{n!}(a+b^n)$ 4(c) $e^{x^2}-e^{y^2}$

6. (a) $\frac{e}{n!}$ (b) 5e 7. (i) e - 2 (ii) $e-\frac{5}{2}$, (iii) $\frac{1}{2}(e-e^{-1})$

1. (i) $\log\left(\frac{5}{4}\right)$ (ii) $\log 3$

9. (i) $138 \cong 14$ 18. $A = B$

19. $y + \frac{y^2}{2} + \frac{y^3}{3} + \ldots.$

21. $x + \frac{x^2}{2} + \frac{x^3}{3} - \frac{3x^4}{4} + \frac{x^5}{5} + \frac{x^6}{6} + \ldots\ldots$

Coefficient of x^{2n+1} is $\frac{1}{2n+1}$ and that of x^{2n} is $-\frac{1}{2n} + (-1)^{n-1}.\frac{1}{n}$

i.e. coefficient of $x^{2n} = \begin{cases} -\frac{3}{2n} & \text{if 'n' is even} \\ \frac{1}{2n} & \text{if 'n' is odd} \end{cases}$

***

(iii) Number of arrangements of n objects, not all different, taken r at a time.

(iv) Number of selection of n objects, not all different, taken all at a time.

(b) Which of the following is not correct ?

(i) ${}^nC_r = {}^nC_{n-r}$

(ii) ${}^nP_r = {}^nP_{n-r}$

(iii) Total number of combinations of n different objects taken some or all at a time is $2^n - 1$.

(iv) Number of selection of four players including the captain out of a team of eleven players is same as the number of selection of three players out of ten players.

(c) Which of the following is the general term in the binomial expansion $(a + x)^n$?

(i) The middle term

(ii) The greatest term

(iii) The r th term

(iv) The (r+1) th term

(d) Which of the following is true for the valid expansion of $(2 + 3x)^{-5}$?

(i) $|x| < \frac{2}{3}$

(ii) $|x| > \frac{2}{3}$

(iii) $|x| < \frac{3}{2}$

(iv) $|x| > \frac{3}{2}$

(iv) $-1 < x \leq 1$

(f) Which of the following is not correct.

(i) The symbol ^nP_r denotes the number of arrangements of n different object taken r at a time.

(ii) $^nC_0 = {}^nC_n$

(iii) Number of ways in which five persons can sit in a row is equal to the number of ways in which they can sit around a round table.

(iv) $^nC_r + {}^nC_{n-r} = {}^{n+1}C_r$

(g) Which of the following is correct ?

The number of ways in which the letters of the word THREE can be arranged in a line is

(i) 5 ! (ii) 5 ! / 3 ! (iii) 5 ! / 2 ! (iv) 3 ! / 2 !

(h) Which of the following is correct ?

Number of five digits numbers that can be formed with the digits 1, 2, 3, 4, 5 in which 3 and 4 come together is

(i) 5 ! (ii) 4 ! (iii) 2 ! (iv) 2 x 4 !

(i) Which of the following is correct ?

The number of ways in which five prizes to be distributed among ten students (with repetition) is

(i) 5 ! (ii) 10 ! (iii) 10^5 (iv) 5^{10}

(i) The total number of terms in the expansion of $(1+x)$ is ________.

(ii) The total number of even numbers which can be formed by using all the digit 0, 1, 3, 4, and 6 is ________.

(iii) The number of ways in which a student can answer five questions out of seven questions is ________.

(iv) In the expansion of $(x-2y)^6$, the general term is ________.

(v) $e^x - e^{-x}$ when expanded in ascending powers of x is ________.

(b) Indicate True (T) or False (F) in the following.

(i) The number of ways in which 3 boys and 4 girls can sit in a row is 3 ! x 4 !

(ii) ${}^nC_r = {}^nC_{n-r}$

(iii) ${}^nP_r = (n-1).{}^{n-1}P_{r-1}$

(iv) The number of ways in which the letters of the word TRIANGLE can be arranged such that the three vowels are together is 8 !

(v) The coefficient of x^6 is the expansion of $(1+x)^{10}$ is ${}^{10}C_6$.

3. **Give short answers to the following questions :**

(a) Distinguish between permutation and combination.

(b) By using factorial notations, show that ${}^{10}C_4 = {}^{10}C_6$.

(c) How many two digits numbers can be formed by using the digits 0, 5, 7 and 8 ?

(d) If ${}^{10}C_k = {}^{10}C_{k+2}$, determine k.

(i) Show that 2. 4. 6. 8........ to n factors is equal to $2^n \times n$!

(j) Expand is $\frac{e^{2x}+1}{e^{x}}$ in ascending powers of x.

ANSWERS

1. (a) (ii) (b) (ii) (c) (iv) (d) (i) (e) (iv) (f) (iv) (g) (iii)
   (h) (iv) (i) (iii) (j) (iv)

2. (a) (i) 2^{10} (ii) 60 (iii) $^7C_5$ (iv) $(-2)^r\,{}^6C_r\,x^{6-r}y^r$ (v) $2(1+\frac{x^2}{2!}+\frac{x^4}{4!}+......)$
   (b) (i) F (ii) T (iii) F (iv) F (v) T

★★★

2.1 INTRODUCTION

Scientific enquiries have always aimed to probe into the mysteries of nature. These enquiries assumed that the behaviour of nature cannot be intricate or complex since our mother 'nature' is benevolent. Thus, the principle of cause and effect dominated the earlier scientific knowledge. This principle is based on laying down sufficient conditions under which a predicted consequence must follow. These were called laws of nature. A stray case of exception was sufficient to invalidiate a law. All laws were rigorous and infalliable. Mathematics as a language, was very suitable to describe.these laws. Some examples of such laws are :- (i) At a constant temperature T, the volume V of a mass of gas is inversely proportional to its pressure P, i.e. PV=C, a constant. (ii) The distance travelled by a rigid body moving with initial velocity u and acceleration f is given by $s = ut + \frac{1}{2}ft^2$ (iii) If a population of size P grows at the rate of r% at the end of every unit of time, its size at the end of t units of time, say P_t, is given by $P_t = P\left(1+\frac{r}{100}\right)^t$.

These are examples of a few situations where the consequence is predictable through mathematical formulae obtained as laws of prediction. Such phenomena are called 'deterministic' and only these were studied in the earlier scientific investigations, thus severely limiting their scope and use.

2.2 PROBABILISTIC PHENOMENA

But there are phenomena which are not predictable even under conditions that seem identical. Some examples are :- (iv) A die or a coin tossed under identical situations can rest on the ground with any one of the faces upwards. It seems meaningless to isolate conditions under which a tossed coin will always show the head side or tail side little is yet known in this regard, inspite of the considerable attention the problem has drawn.

The category of phenomena from which the above examples are samples are 'Probabilistic'. Probabilistic phenomena are not predictable . But for many of them, the degree of unpredictability can be measured under very general assumptions. A precise knowledge of this degree of unpredictability can be of considerable practical use. For example, in the toss of a coin, assuming the coin as 'fair', we may describe the outcome of any future toss as a head or a tail with 50% chance or probability of each. Obviously, this is precisely the best answer one can give about any future toss. In the language of probability we say that head and tail are equally likely or probable and the probability of each is 0.5 or 50%.

2.3 HISTORICAL BACKGROUND

The idea of measuring probability or the chance of occurrence of an uncertain event is as old as gambling with dice. Pascal and Fermat, two French mathematicians of the seventeenth century formalised the concept of probability in some problems of gambling. The earliest publications on probability theory are "Treatise on Probability" by Jacob Bernoulli (1654-1705), posthumously published in 1713 and "Doctrine of Chance" of De Moivre (1667-1754) published in 1718. Notable among other important workers in the subject in the nineteenth century were P.S. Laplace and T. Bayes. The Russian School of Mathematicians have made very important contributions to the theory of probability. Notable among them were Chebyshev (1821-1894), Markov (1856-1922), Liapounov, and Khintchine. The foundations of modern probability theory were laid down by Von Mises, A.N. Kolmogorov, Paul Levy and William Feller. Presently, almost all branches of applied sciences use probability as an important tool to create knowledge.

scores. The probability (1/6) may be assigned to each outcome. Also, while drawing a card from a well-shuffled full deck, the outcome set has 52 elements, each element representing one card. A probability (1/52) may be associated with each element, since the deck is well-shuffled. These situations may be generalised as a model which is described below-

(1) **Trial :** A trial (or an experiment) is the creation of identical situations. Tossing of a die, or drawing of a card from a full deck are examples of trials. We consider trials, the outcomes of which are uncertain.

(2) **Elementary Event :** Every possible outcome of a trial is called an elementary event. In tossing a die, appearance of any score like 5 or 6 is an elementary event. In drawing a card, draw of a spade ace or three of diamond is an elementary event. There are 6 elementary events in the toss of a die, and 52 elementary events in the draw of a card from a full deck.

(3) **Event :** Any defining property may specify an event. For example, in tossing a die, appearance of an even number is an event. This event occurs if 2, or 4, or 6 appears in a toss. Thus, it consists of three elementary events. Similarly, in draw of a card from a full deck, appearance of a spade or a picture card is an event. The former consists of 13 elementary events as there are 13 cards of spades in the deck. The defining property dichotomises the elementary events into two non-overlapping categories - some favourable to the event and others unfavourable.

(4) **Exclusive Events :** Two events are said to be exclusive if both of them cannot happen simultaneously. For example, in drawing a card from a full deck, appearance of a heart and a spade are exclusive events. Appearance of a picture card and

(5) **Exhaustive events :** A set of events is exhaustive, if every elementary event is included in one or more of these events. This means, every possible outcome will result in the happening of at least one of these events. For example, in the selection of a number from the set {1, 2, 3, 4,.....,17, 18, 19, 20}, the events (i) selection of an even number (E_1), (ii) selection of a single digit number (E_2), (iii) selection of a prime number(E_3), and (iv) selection of a multiple of 5 (E_4); are mutually exhaustive.

(6) **Mutually exclusive and exhaustive events :** A set of events are mutually exclusive and exhaustive if (i) they are exclusive events and (ii) they are exhaustive events. For example, in selecting a number from {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, the events $E_1 = \{1, 4, 7, 10\}$, $E_2 = \{2, 5, 8\}$, and $E_3 = \{3, 6, 9\}$ are mutually exclusive and exhaustive events.

(7) **Equally Likely events :** A set of events are equally likely if the conditions of the trial do not favour any one of them in preference to any other event of the set. For example, in tossing a homogeneous die, having the shape of a cube, any one of the six faces is as likely to show up as any other. Similarly, in choosing a card from a well-shuffled deck, any card is as likely to appear as any other card. But the appearance of a spade and a king are not equally likely as there are more spades than kings. Hence, a spade is more likely to be selected than a king.

Classical Probability : If the outcome of a trial is any one of n equally likely events, then a total probability of 1 (or unity) is distributed equally among these events. Hence, the probability of any elementary event is (1/n). Here, 1 represents the certain event that one of these elementary events must happen in each trial. We associate '0' (Zero) probability

$$P(E) = \frac{m}{n} = \frac{\text{number of favourable cases}}{\text{number of possible cases}}$$

By the same definition, the probability of E not happening $(\bar{E})$ is -

$$P(\bar{E}) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - P(E)$$

$$\text{Thus, } P(E) + P(\bar{E}) = \frac{m}{n} + 1 - \frac{m}{n} = 1$$

The above definition cannot be used to compute probability of an event, unless the outcomes are equally likely cases. For purposes of illustration, we solve some problems of computing probabilities of events, where the trial is designed or manipulated to make the outcomes equally likely. Such trials are called random trials or random experiments.

2.5 ILLUSTRATIVE EXAMPLES

Example 2.1 : Two dice are simultaneously tossed. What is the probability of getting a total score which exceeds 10 ?

Solution - Let the dice be numbered as 1 and 2. Let x be the score obtained in die 1 and y the score on die 2. The possible values of (x, y) such that $(x + y) > 10$ are x=5, y=6 ; x=6, y=5; and x=6, y=6. This gives 3 favourable pairs. The possible number of pairs (x, y) are $6 \times 6 = 36$, since x and y can take 6 values each. Hence

$$\text{Required probability} = \frac{3}{36} = \frac{1}{12}$$

Eaxmple 2.2 : Three numbers are randomly drawn from the set {1, 2, 3, 4, ..., 12, 13, 14, 15}. What is the probability that their sum will be an odd number ?–

Number of ways of selecting 2 even and one odd number = ${}^{7}C_2 \times 8 = 21 \times 8 = 168$

$$\therefore \text{ Required Probability} = \frac{56+168}{455} = \frac{224}{455} = \frac{32}{65}$$

Example 2.3 : What is the probability that the days of the week on which three persons selected at random will be born will be (i) all different (ii) not all same ?

Solution : Let the 3 persons be named A, B, C and their birthdays of the week be recorded that order.

Hence, the total number of possible orderings = $7 \times 7 \times 7 = 7^3$

(i) Number of possible orderings where the week days are all different = $7 \times 6 \times 5$, since the days already recorded have to be excluded in the next recording.

$$\therefore \text{ Required probability} = \frac{7 \times 6 \times 5}{7^3} = \frac{30}{49}$$

(ii) Number of possible orderings where all days are same = 7, since the same days have to be repeated.

So, number of orderings where not all days will be same = $(7^3 - 7)$

$$\therefore \text{ Required probability} = \frac{7^3 - 7}{7^3} = 1 - \frac{1}{49} = \frac{48}{49}$$

Example 2.4 : A pack of 4n cards containing 2n red and 2n black cards is randomly divided into two halves. Find the probability that there will be n red and n black cards in each half.

Example 2.5 : The numbers {1, 2, 3, 4, 5, 6, 7, 8} are randomly arranged in a sequence. What is the probability that 1 and 3 do not occupy consecutive places ?

Solution : Total number of possible arrangements = 8!

If 1 and 3 are considered one unit, number of arrangements in which 1 and 3 occupy consecutive place = 2 ×7!, since 1 and 3 can occur together as '13' or '31'.

So, Probability that 1 and 3 occur together $= \frac{2 \times 7!}{8!} = \frac{1}{4}$

$\therefore$ Probability that 1 and 3 do not occupy consecutive places $= 1 - \frac{1}{4} = \frac{3}{4}$

Example 2.6 : Three numbers are randomly selected from the set {1, 2, 3, ..., 22, 23, 24, 25} . What is the probability that there will be (i) exactly 2 odd numbers (ii) at least one odd number, in the choice ?

Solution : Number of possible ways of selecting 3 numbers out of the 25 numbers in the set $= {}^{25}C_3$

(i) Number of ways of selecting 2 odd numbers from the 13 odd numbers in the set $= {}^{13}C_2$

$\therefore$ Number of ways of selecting 2 odd and one even number from the 12 even numbers in the set $= 12 \times {}^{13}C_2$

So, probability that there will be no odd number in the choice = $\frac{C_3}{^{25}C_3} = \frac{11}{115}$

$\therefore$ Required probability of at least one odd number being included in the choice

$= 1 - \frac{11}{115} = \frac{104}{115}$

Example 2.7 : Five numbers are randomly selected from the set {1, 2, 3, 4, ..., 27, 28, 29, 30}. Find the probability that the third smallest selected number is 10.

Solution : Number of ways of selecting 5 numbers from 30 = $^{30}C_5$. The third smallest number is 10, when 2 of the numbers are from the set {1, 2, 3, ..., 7, 8, 9} and 2 from the set {11, 12, 13, ..., 28, 29, 30} and the fifth number is 10.

So, number of favourable ways = $^{9}C_2 \times {}^{20}C_2$

$\therefore$ Required probability = $\frac{^{9}C_2 \times {}^{20}C_2}{^{30}C_5} = \frac{380}{7917}$

Example 2.8 : Five 'a' s and eight 'b' s are written in a sequence of 13 letters. Find the probability that no two 'a' s occupy consecutive places.

Solution : Total number of possible sequences in which five 'a' s and eight 'b' s can be written = $^{(8+5)}C_5 = {}^{13}C_5$

So, number of favourable sequences = ${}^{9}C_5$

$$\text{Required Probability} = \frac{{}^{9}C_5}{{}^{13}C_5} = \frac{14}{143}$$

Example 2.9 : Three numbers are randomly selected from the (2n+ 1) natural numbers {1, 2, 3, ..., (2n -1), 2n, (2n + 1)}. What is the probability that they will be in arithmetic progression (A.P) ?

Solution : Number of ways in which 3 numbers can be selected from the (2n + 1) numbers

$$= {}^{(2n+1)}C_3 = \frac{n(4n^2-1)}{3}$$

(i) Number of triplets in A.P. with common difference d=1 are (1, 2, 3), (2, 3, 4), (3, 4, 5),...., (2n – 1, 2n, 2n + 1) which total to (2n – 1) triplets.

Number of triplets with d=2 are (1, 3, 5), (2, 4, 6), ..., (2n-3, 2n-1, 2n+1) which total to (2n-3) triplets. Similarly, there are (2n-5) triplets with d=3, (2n-7) triplets with d=4, and so on untill the last one triplet with d=n which is (1, n+1, 2n+1)

Hence, the total number of possible triplets in A.P. = (2n – 1) + (2n – 3) + (2n – 5) + ... +3 + 1= n^2.

$$\text{Required probability} = \frac{3n^2}{n(4n^2-1)} = \frac{3n}{4n^2-1}$$

$x^\alpha \times x^\beta \times x^\gamma = x^{\alpha+\beta+\gamma}$. This gives a 1-1 correspondance between (α, β, γ) and $x^{\alpha+\beta+\gamma} = x^\alpha \times x^\beta \times x^\gamma$. Now,

Consider the product$(x+x^2+x^3+x^4+x^5+x^6)$ $(x+x^2+x^3+x^4+x^5+x^6)$ $(x+x^2+x^3+x^4+x^5+x^6)$. When the above 3 factors are multiplied we get all terms of the type $x^\alpha \times x^\beta \times x^\gamma$, where x^α is selected from the first factor. This means a score of α in the first toss. x^β and x^γ have similar meanings. The number of favourable cases correspond to possible number of (α, β, γ) triplets with $\alpha+\beta+\gamma=12$. Hence, out of a total of $6\times6\times6=216$ terms of the type $x^\alpha \times x^\beta \times x^\gamma$, the favourable cases are given by the coefficient of x^{12} in the above product. This coefficient counts the number of triplets (α, β, γ) with $\alpha+\beta+\gamma = 12$

Now $\quad (x+x^2+x^3+x^4+x^5+x^6) = x(1+x+x^2+x^3+x^4+x^5)$

$$= x(1-x^6)(1-x)^{-1}$$

Hence, the product $(x+x^2+x^3+x^4+x^5+x^6)^3 = x^3(1-x^6)^3(1-x)^{-3}$

Expanding the R.H.S by binomial theorem, we get,

L.H.S. $= x^3(1-3x^6+3x^{12}-x^{18})$

$$\times\left(1+3x+\frac{3x4}{2}x^2+\frac{4x5}{2}x^3+\ldots+\ldots+\frac{(n+1)(n+2)}{2}x^n+\ldots\right)$$

Coefficient of x^{12} in the R.H.S

Alternative Method

Let x, y, z be the scores obtained in the 3 tosses. Each of the three numbers x, y, z can take the possible values 1, 2, 3, 4, 5, 6, So,

Total number of possible triplets (x, y, z) = $6 \times 6 \times 6 = 216$

To count the number of favourable cases where $x + y + z = 12$, we can use direct enumeration as follows :

(i) All the numbers x, y, z are equal to 4.

No of such cases = 1

(ii) Two of the numbers are equal and the third is different. This can occur if the such of values are :- 3, 3, 6; 5, 5, 2. Each of the set of values can be permuted to give (x, y, z)

Henc, the number of such cases = $\frac{3!}{2!} + \frac{3!}{2!} = 6$

(iii) All the 3 numbers are different. This can happen if the set of numbers are :- 1, 5, 6; 2, 4, 6; and 3, 4, 5. Each one of these sets can be permuted in 3! ways.

Hence the number of such cases = $3 \times 3! = 18$

Thus the total number of favourable cases = 1 + 6 + 18 = 25

∴ Hence, the required probability = $\frac{25}{216}$

3. In one toss of a die, give examples of (i) two exclusive events (ii) two exhaustive events and (iii) three events which are mutually exclusive and exhaustive.

4. Two cards are randomly drawn from a well-shuffled pack. Find the probability that (i) they are of the same denomination (ii) they are of the same suit.

5. Two dice are thrown. Find the probability that the difference between the two scores obtained is an even number, which includes zero.

6. In a single throw of two dice, which total score has the least probability ?

7. Both Ram and Sam throw a die each. Find the probability that (i) they throw the same scores (ii) Ram throws the same score as Sam, when Sam is known to have thrown a 3 score.

8. What is the probability that a randomly selected leap year has (i) 53 Sundays (ii) 53 Saturdays and 53 Sundays.

9. The letters of the word AUTOMOBILE are arranged in a random sequence. What is the probability that the two O's are separated by a single letter ?

10. The digits 0, 1, 2, 3, 4, 5, 6, 7, 8 are arranged at random. Find the probability that the number obtained is (i) divisible by 5 (ii) an even number exceeding 10^8.

11. An urn contains 4 red, 5 black, and 1 blue ball. Three balls are chosen at random. Find the probability that the blue ball will be included in the choice.

12. Four numbers are randomly selected from the set {1, 2, 3, 4 ..., 15, 16, 17}. Find the probability that at least one multiple of 3 will be included in the choice.

13. Two numbers are randomly selected from the set {1, 2, 3, ..., 18, 19, 20}. Find the probability that the sum is divisible by 5.

17. 20 chairs have been arranged in a row to accommodate 12 boys and 8 girls. If they occupy the seats at random, what is the probability that no two girls occupy consecutive chairs ?

18. Four numbers are successively drawn from the set {1, 2, 3, 4, 5} replacing the number drawn in every draw before the next draw. Find the probability that the sum of the numbers is 15.

2.6 LIMITATIONS OF CLASSICAL PROBABILITY

(1) The definition of classical probability cannot be applied to probabilistic situations, where the outcomes are known but not equally likely. For example, an agriculturist or a meteorologist will not be able to associate probabilities with good or moderate or poor rainfall in the next monsoon season due to arrive shortly, using the classical definition. These probabilities are not equally likely and thus the classical definition is not applicable. Similarly, the probability that Ram, who is aged 45 years as on to-day, will be alive after 10 years cannot be asserted as 50%; since the two possibilities that he may or may not be alive are not equally likely. (2) In some probabilistic situations, a variable may take an infinite number of values. For example, the height or life of a person can take any value between two possible limits. The interval has infinite number of values of the variable height. So, the classical definition breaks down. This led to the concept of empirical or statistical probability which is defined below.

Statistical or Empirical Probability - Definition (Due to Von Mises) - Suppose a trial is repeated n times under similar or identical situations. If an event E happens in m out of these n trials and does not happen in the rest, then (m/n) is called the observed relative frequency of E. Von Mises defined the probability of E, P(E), as the limiting value of (m/n) as n tends to infinity. Thus, science. Hence, both a sound theoretical base as well as precise computation were important requirements. This culminated in the modern axiomatic approach to probability which unified the classical and empirical approaches and generalised it to a broader field of application. Before describing the axiomatic approach, we describe its mathematical pre-requisites.

2.7 MATHEMATICAL PRELIMINARIES

Set- A set is a well-defined aggregate or collection of objects. Examples of sets are :- (i) All students of a class (ii) All positive integers ≤ 1000 (iii) All points on a line segment or ray (iv) All equilateral triangles (v) The 26 letters of the English alphabet (vi) All voters of an electorate (vii) All animals of a zoo etc.

Elements : The smallest individual units which build a set are called its elements. For example, a student identified by his roll number, an equilateral triangle with each side of length 6 cms, a particular fish in the zoo aquarium are elements of the corresponding sets.

Universal Set : In all our discussions, there is a maximal set which limits the scope of our discourse. This is called the Universal Set. All our statements apply within the premises of this Universal set or Universe; usually denoted by the symbols U, S or W. In tossing of a coin U={H, T}, a set with two elements Head (H), and Tail (T). In tossing a die U={1, 2, 3, 4, 5, 6} is a small part of the much bigger U={1, 2, 3, ...98, 99, 100}, which is the universal set of all positive integers upto 100. Obviously, in tossing a coin possibilities like the coin standing on an edge or rolling away are not within the universe of our discourse. While describing a finite set, we may write all the elements inside curly brackets like {1, 2, 3, 4} or write a generic symbol and then indicate a property defining the set. For example $\{x: x^2 \leq 1\}$ includes all numbers x, such that $x^2 \leq 1$.

$A \subseteq B$ and $B \subseteq A \Leftrightarrow A = B$.

Operations on sets

Union : The union of A and B, written as $A \bigcup B$ and read as 'A union B' is a set that contains all elements included in A or B or both these sets.

Thus,

$A \bigcup B = \{x : x$ belongs to at least one of the two sets and A and B$\}$. Similary, we define the union of n sets $A_1, A_2, \ldots, A_n$ as

$$\bigcup_{i=1}^{n} A_i = \{x : x \text{ belongs to at least one of the sets } A_1, A_2, \ldots, A_n\}$$

Intersection : The intersection of A and B, written as $A \bigcap B$ and read as 'A intersection B' is a set that contains all elements included in both these sets. Thus,

$$A \bigcap B = \{x : x \in A \text{ and } x \in B\}$$

If no element is common to A and B, we say A and B are disjoint or exclusive and write $A \bigcap B = \phi$, where ϕ called the 'empty set' or 'null set' is a set with no elements.

Similarly, the intersection of n sets $A_1, A_2, \ldots, A_n$ is the set containing all elements which are included in each of the A_i 's i = 1, 2, 3, ..., n. Thus,

$$\bigcap_{i=1}^{n} A_i = \{x : x \in A_i \text{ for } i = 1, 2, 3, \ldots, n\}.$$

Complement- The complement of a set A, denoted by $\overline{A}$ or A^c is the set which includes all elements not included in A. Thus,

$$A^c = \{x : x \notin A\}$$

Obviously $U^c = \phi$ and $\phi^c = U$

(i) $A \cup A^c = U, A \cap A^c = \phi, U - A = A^c, A \cup U = U$,

$A \cup \phi = A, A \cap \phi = \phi, A \cap U = A, (A^c)^c = A$

The above operations also satisfy the following laws.

Commutative laws : $A \cup B = B \cup A, A \cap B = B \cap A$

Associative laws : $(A \cup B) \cup C = A \cup (B \cup C)$

$(A \cap B) \cap C = A \cap (B \cap C)$

Distributive laws : $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Idempotency Laws : $A \cup A = A, A \cap A = A$

De Morgan's Laws : $\left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c$

$$\left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c$$

Remark : De Morgan's laws also hold for a countable number of events $A_1, A_2, A_3, \ldots$ i.e. when n tends to infinity.

Venn Diagrams - A geometrical representation of an event in the form of the interior of a closed curve is often helpful in understanding these operations discussed above. In Venn diagrams, the universal set may be represented by a larger closed area in a plane. The events may be represented by smaller closed curves entirely inside the universal set. In the Venn diagram shown here, the large rectangle is the Universal set U. A and B are represented by two circular regions i.e. the areas inside the circles. Region 1 represents solving any problem.

Sequence of sets : Limits

Consider a sequence of sets $A_1, A_2, A_3, ...$from U, which is usually written as $\{A_n\}$. We define,

(i) **Limit superior of $\{A_n\}$**- It is a set which includes all elements which are present in infinitely many of the sets in the sequence. In symbols-

Lim sup $A_n = \{x : x \in A_n$ for an infinite number of values of n}

(ii) **Limit Inferior of $\{A_n\}$**- It is a set which includes all elements which are present in all except a finite number of sets in the sequence. In symbols-

Lim inf $A_n = \{ x : x \in A_n$ for all but a finite number of values of n}

Obviously Lim Inf A_n is a subset of Lim sup A_n. If,

Lim sup A_n = Lim inf A_n = A (Say)

we say the sequence A_n has a limit equal to A, written as,

Lim A_n = A

In the following cases, the sequence A_n has a limit :-

(i) If an $A_n \subseteq A_{n+1}$ for all n = 1, 2, 3, ... ; we say such a sequence of sets is **monotonic increasing**. Such sequences have a limit.

(ii) If $A_{n+1} \subseteq A_n$ for all n = 1, 2, 3, ... ; we say such a sequence of sets is **monotonic decreasing**. Such sequences have a limit.

A monotonic increasing or monotonic decreasing sequence is commonly called a monotone sequence.

Examples :

(i) Suppose U = {1, 2, 3} and $A_n = \{a_{n1}, a_{n2}\}$, where a_{n1} and a_{n2} represent two sequence in A_N will continue to remain in all subsequent A_i's i.e at all subsequent stages.

(iii) Suppose $U = \{x : 0 \le x \le 1\}$ and $A_n = \left[0, \frac{1}{n}\right]$.

This sequence is monotonic decreasing and Lim $A_n = \{0\}$ i.e. a set which includes only '0'. But if $B_n = \left(0, \frac{1}{n}\right)$ then Lim $B_n = \phi$.

2.8. CLASS OF SETS

A class of sets is an aggregate of several sets which is well-defined. The following two classes of sets are used to build a probability structure which assigns a probability to each set of the class.

Algebra of Sets - A class of sets F is called an algebra if

(i) $A \in F, B \in F \Rightarrow A \bigcup B \in F$ and (ii) $A \in F \Rightarrow A^c \in F$

Corollary- If $A_1, A_2, \ldots, A_n \in F$, then $\bigcup_{i=1}^{n} A_i \in F$.

Proof :- This can be proved by induction on n. Suppose, the statement holds for (n-1) i.e.

$A_1, A_2, \ldots, A_{n-1} \in F \Rightarrow \bigcup_{i=1}^{n-1} A_i \in F$.

Then $= \bigcup_{i=1}^{n} A_i = \left(\bigcup_{i=1}^{n-1} A_i\right) \bigcup A_n \in F$ by property (i) given above. Hence, the corollary is true for n, if it is true for (n-1). We know it is true for n = 2. Hence, by the principle of induction it is true for all n = 2, 3, 4, ...

$_1,\ _2,\ _3, \qquad \bigcup_{i=1}$

The above properties state that countable union and complements of elements already in **F** give elements which are already in **F**. We thus, say that a σ - algebra is closed under countable union and complementation. σ - algebra is also called σ - filed.

Some Properties - An algebra is also closed under most of the other operations described above. We describe some of these properties below and indicate their proofs.

(a) An algebra is closed under finite intersection.

Proof :- By De Morgan's laws, $\left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c$

$$\text{So, } \left(\bigcap_{i=1}^{n} A_i\right) = \left(\bigcup_{i=1}^{n} A_i^c\right)^c$$

The R H S of the above equation belongs to **F** by the properties (i) and (ii) of an algebra.

(b) $A \in F, B \in F \Rightarrow (A - B) \in F$.

Proof : $(A - B) =, A \cap B^c$ which belong to F by (a)

(c) U and ϕ belong to every algebra.

Proof : $A \in F \Rightarrow A - A = \phi \in F$

$A \in F \Rightarrow A^c \in F$. Hence, $A \cup A^c = U \in F$

(d) Every σ - algebra is also an algebra.

Proof : $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{\infty} A_i$ where $A_{n+1}, A_{n+2}, A_{n+3}, \ldots$ are all equal to ϕ, the empty set which belongs to F. Hence the statement of (d) follows :

Our subsequent discussions will mostly centre on algebras.

Exercise - 2 (B)

1. If U be the set of all non-negative integers, write down all the elements of the following sets-

   (a) All prime numbers ≤ 20

   (b) All perfect squares ≤ 50

   (c) All multiples of 5 in the closed interval [20, 40].

   (d) All numbers in the interval [20, 50] which are diviscible by 7.

   (e) All numbers ≤ 21 which are the products of exactly three prime factors.

2. Set U = [0, 1]. If A=[0, 0.4], B=[0.6, 0.8], C={x : x $\geq$ 0.5}; write down the following sets (i) $B \cup C$ (ii) C -B (iii) $(A \cup B \cup C)^c$

3. U Consists of all numbers of the form (m/n) where m and n are non-negative integers such that m+n = 20. Write down the following subsets of U. (i) All positive integers (ii) All numbers which can be represented as a terminating decimal. (iii) All numbers which exceed 1.5.

4. U = {a, b, c, d, e, f, g, h, i, j}. If A = {a, b, c, d}, B = {a, e, i, j}, C= {c, e, h, i} and D ={f, g, h, i, j}; write down the following sets. (i) $\overline{A \cup D}$ (ii) $A \cap B \cap C$ (iii) $(A \cup B) \cap D$ (iv) B - C.

5. Prove that if F_1 and F_2 are two algebras in U, then $F_1 \cap F_2$ is also an algebra in U.

2.9 AXIOMATIC DEFINITION OF PROBABILITY

We have indicated earlier that the definition of probability needs to be extended to probabilistic situations beyond the limited few where there are equally likely outcomes or where repeated trials under similar conditions can be conceived. This led to the

Random Experiment - An experiment is done by creating some predetermined conditions. The experiment is random when its ·outcome is uncertain. Apart from tossing a die or a coin, selecting a unit from a list by lottery, observing the response of a patient under specific medication, or measuring the yield of a plant under a fertilizer treatment are also examples of random experiments.

Trial - A single experiement is a trial

Outcome - Each possible consequence of a trial is an outcome. We have some possible outcomes out of which one outcome materialises. In an agricultural experiment an outcome may mean the yield of a single plant or a plot. In a lottery, an outcome is chance selection of a single unit.

We denote a random experiment as E and the possible outcomes as S= $\{e_1, e_2, e_3, \ldots, e_m\}$ m being the number of possible outcomes. A trial results in a single outcome, say e_i. Each e_i is also called an elementary event. In some trials, the possible outcomes may not be finite but a countable sequence.

Sample Space- The set of all possible outcomes S is called the sample space. The number of elements in S need not always be finite. For example, an outcome may be a measurement on a continuous scale. But, we will mostly confine our discussions to sample spaces with a finite number of outcomes.

Event- An event is characterised by its description or by a defining property. A, B, C etc. are the common notations for an event.

Some elementary events are favourable to A i.e. A happens only if one of these elementary events happens. Others are unfavourable to A i.e. if any one of them happens, A does not happen. For example, in throwing of a die, getting an even score is an event which happens if the score is 2, or 4, or 6. Birth of a male child is an event in recording the sex of a new born baby.

Example 1- A coin is tossed three times and the outcome noted for each toss as H(Head) or T (Tail). The sample space is -

S= {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

There are eight elements or elementary events of S. Examples of some events are (i) A, the event that there are more heads than tails. A = {HHT, HTH, THH, HHH} (ii) B the event that there are exactly two tails, B= {TTH, THT, HTT}. (iii) C the event that only one side show up in all 3 tosses, C = {HHH, TTT}

Example 2- A die is tossed twice and the scores recorded for both the tosses The sample space S is -

S={(1, 1)}, (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

S consists of 6 × 6 = 36 elementary events. Examples of events are :-

(i) A, the sum of the scores is 8 = {(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)}

(ii) B, the sum of the scores is less than 4= {(1, 1), (1, 2), (2,1)}

(iii) C, the difference between the two scores is zero

= {(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)}

Example 3 - Two numbers are randomly drawn from the set {1, 2, 3, 4, 5} and the numbers recorded in ascending order, disregarding the observation sequence. The sample space is

S = {(1, 2), (1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (3, 4), (3, 5), (4, 5)}

Both the numbers can be drawn simultaneously or in succession, where the number drawn first is not replaced before the next draw. *S* has 10 elementary events.

F stand for a male child and a female child respectively. Here, the families having both boys and girls are the subset A = {MF, FM} and A^c stands for families with only boys or only girls.

Remark - The events described above have their defining properties. But any collection of elementary events is an 'event'.

Analogy between subsets and events - In our earlier discussion. We have described an algebra and a σ-algebra of a class of sets from a universal set. The concepts given there can be directly extended to events of a sample space. We briefly describe this extension below :-

Exclusive events - Events A and B are exclusive if they cannot happen together i.e. if one event excludes the other. Hence, $e \in A \Rightarrow e \notin B$. A set of events are mutually exclusive if they are exclusive pairwise i.e. no two of them can happen together.

Exclusive events are also called **disjoint events.**

Exhaustive events - A set of events are (mutually) exhaustive if every elementary event e belongs to one or more of these events.

Mutually exclusive and exhaustive events - Events $A_1, A_2, \ldots, A_k$ are mutually exclusive and exhaustive if they are pairwise exclusive and if every elementary event belongs to one of these events. Obviously no elementary event can occur in two of them. These concepts have already been discussed earlier.

The operations defined for a class of sets can also be extended to a class of events. This is described below -

Union $\bigcup$ - If A and B are two events, C = $A \bigcup B$ defines as event C which happens if at least one of the two events A and B happens. It means the happening of A or B or both.

included in more than one A_i's are included in this countable union.

Intersection $\bigcap$ - If A and B are any two events, $C = A \bigcap B$ defines an event C which happens if both A and B happen. If A and B are exclusive, the intersection is the impossible event ϕ. Similarly, $\left(\bigcap_{i=1}^{n} A_i\right)$ is an event which happens if each one of the n events $A_1, A_2, ..., A_n$ happens. A similar definition holds for the intersection $\left(\bigcap_{i=1}^{\infty} A_i\right)$ of a countable sequence of events $A_1, A_2, A_3...$. Sometimes, we drop the symbol $\bigcap$ and write $A \bigcap B$ and AB synonymously. ABC has a similar meaning of the happening of A, B and C.

Complement - Complement of an event A is an event which includes all elementary events not in A. Denoted by A^c or $\overline{A}$ it happens when A does not happen.

Hence, $A^c = \{e : e \notin A\}$

If A^c happens, A does not happen. The converse is also true.

Difference - For any two events A and B, C = (A - B) is an event which happens if (i) A happens but (ii) B does not happen.

A implies B - If all elementary events favourable to A are also favourable to B, we say A implies B. Thus B happens whenever A happens. Symbolically we write $A \subseteq B$ i.e. A is a subset of B. If A implies B and B implies A, then A and B are the same events. $A \subseteq B$ is also written as $B \supseteq A$ which may be read as "B is implied by A".

The usual laws of sets like the commutative laws, associative and distributive laws, Do Morgan's laws etc. also hold for a set of events.

(ii) Exactly one of the two events happens can be written as $(A-B)\bigcup(B-A)$ or $AB^c \bigcup BA^c$.

(iii) Neither A nor B happens can be written as $(A\bigcup B)^c = A^c \bigcap B^c$.

Example 2 - Suppose we have A, B, C i.e. three events under consideration. Each one may or may not happen.

Then

(i) A and B only happen can be written as ABC^C.

(ii) Exactly two of the 3 events happen can be written as $ABC^C \bigcup AB^CC \bigcup A^CBC$.

(iii) None of the three events happens can be written as $A^cB^cC^c$. Similarly,

(i) $(ABC \bigcup A^cB^cC^c)^c$ means at least one but not all the three events A, B, C happen.

(ii) $(A \bigcup A^cB \bigcup A^cB^cC)$ means at least one of the three events A, B and C happens. This event is the same as $(A \bigcup B \bigcup C)$.

2.10 ALGEBRA OF EVENTS

A class of events is called an algebra if the corresponding class of sets is an algebra. Similarly, a class of events is called a $\sigma -$ algebra (sigma algebra) if the corresponding class of sets is a $\sigma -$algebra.

An algebra or a $\sigma -$algebra of events is logically complete with respect to some operations defined earlier. If a given class of events (sets) is not an algebra or $\sigma -$ algebra; then it can be augmented by including the extra events that arise as a result of these operations and construct a smallest algebra or $\sigma -$algebra which includes the given class of events (sets).

closed under these operations. We now proceed to define probability on the events of an algebra of events.

Probability - Most generally, probability is defined on the events of an algebra. The definition is axiomatic i.e. it follows certain axioms or assumptions which are fundamental. These axioms cannot be questioned and help the construction of a probability structure. This model of a probability distribution is not obtained from any individual or group of problems. But one of these models can be used to describe any real probabilistic problem. Probability is defined on the elements of an algebra of events, which are events or sets. Hence it is a set function. For an algebra **F**, the probability of an event A, denoted by P(A), is a non- negative, finitely additive set function such that P(S) = 1. Thus, P(A) satisfies the conditions-

(i) $0 \le P(A) \le 1$

(ii) If $A_1, A_2, \ldots, A_n$ are mutually exclusive events, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

(ADDITIVITY)

(iii) If S is the sample space, then P(S) = 1.

Such a set function is called a **probability measure.** It may be noted that the classical definition satisfies these properties. The following important results follow from these conditions.

(i) Since $S \cap \phi = \phi$ and $S \cup \phi = S$, $P(S \cup \phi) = P(S) + P(\phi) = P(S)$ which implies $P(\phi) = 0$

(ii) Since, $A \cup A^c = S$, and $A \cap A^c = \phi$, $P(A \cup A^c) = P(A) + P(A^c) = P(S) = 1$,
$\therefore P(A^c) = 1 - P(A)$

A^c , S}. A probability measure here is defined by P(A) = p, $(0 \leq p \leq 1)$. The probabilities of other events are determined by the axioms. Thus, $P(\phi) = 0$, P(S) = 1, and $P(A^c) = 1-p$. For a particular situation, p may be determined by the mechanism of the experiment, like $p = \frac{1}{2}$ when a coin is tossed and A is the appearance of a head. It may also be approximated by using Von Mises concept of relative frequency, or by any intelligent guess, or even observation by experience.

Extension of Probability Measure - A probability measure P can also be defined on a σ-algebra of events. Here, the condition (ii) above is modified to a countable sequence of disjoint (exclusive) events. Thus, P(A) satisfies

(i) $0 \leq P(A) \leq 1$

(ii) If $A_1, A_2, A_3 \ldots$ be a countable sequence of disjoint events, then,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

(iii) If S is the sample space, then P(S) = 1.

This concept is necessary if probability has to be defined on each of a countable sequence of disjoint events. For example, if a coin is tossed until the first head is obtained, the tosses will terminate with n tosses if (n-1) tails are followed by the first head. Here,

$$P \text{ (n tosses)} = P(E_n) = \frac{1}{2^n}, n=1, 2, 3\ldots$$

The number of outcomes here is infinity and all of them are elements of a σ – algebra, A σ – algebra is the smallest closed class of sets which included $E_1, E_2, E_3, \ldots$ as elements .

Also A and A^cB are disjoint and $A \cup A^cB = A \cup B$,

Hence, $P(A \cup B) = P(A) + P(A^cB)$ (from 2.1)

$= P(A) + P(B) - P(AB)$

as required.

Corollary 2.1.1 - $P(A \cup B) \leq P(A) + P(B)$

This follows from theorem 2.1 as $P(AB) \geq 0$

Corollary 2.1.2 - For a set of events $A_1, A_2, \ldots, A_n$,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

Proof : The result easily follows by induction.

Suppose the result holds for (n - 1) events i.e.

$$P\left(\bigcup_{i=1}^{n-1} A_i\right) \leq \sum_{i=1}^{n-1} P(A_i)$$

Then,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P\left[\left(\sum_{i=1}^{n-1} A_i\right) \cup A_n\right]$$

$$\leq P\left(\bigcup_{i=1}^{n-1} A_i\right) + P(A_n)$$

Corollary 2.1.3 - $P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$

Proof :- By corollary 2. 1. 2,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

or $$1 - P\left(\bigcup_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(A_i)$$

or $$P\left(\bigcup_{i=1}^{n} A_i\right)^c \geq 1 - \sum_{i=1}^{n} \left[1 - P(A_i^c)\right]$$

or $$P\left(\bigcap_{i=1}^{n} A_i^c\right) \geq 1 - n + \sum_{i=1}^{n} P\left(A_i^c\right) \qquad \text{(De Morgan's law)}$$

$$= \sum_{i=1}^{n} P\left(A_i^c\right) - (n-1)$$

Theorem 2.2 - For three events A, B, C,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

Proof : $P(A \cup B \cup C) = P\left[(A \cup B) \cup C\right]$

$$= P\left[(A \cup B)\right] + P(C) - P\left[(A \cup B) \cap C\right] \qquad \text{(By Theorem 2.1)}$$

$$= P(A) + P(B) - P(AB) + P(C) - P\left[(AC) \cup (BC)\right] \qquad \text{(Distributive law)}$$

$$= P(A) + P(B) + P(C) - P(AB) - \left[P(AC) + P(BC) - P(ABC)\right] \qquad \text{(By Theorem 2.1)}$$

Solution : Let the two dice be numbered as 1 and 2. Let the score on die 1 be x and the score on die 2 be y. The sum of the scores $S = (x + y)$. Now, the event $S \leq 4$ can occur in the following exclusive ways :- (i) $S = 2$ (ii) $S = 3$ (iii) $S = 4$

So, $P(S \leq 4) = P(S = 2) + P(S = 3) + P(S = 4)$

Sipce both x and y can take any values from the set {1, 2, 3, 4, 5, 6} the pair (x, y) has $6 \times 6 = 36$ possibilities, all equally likely.

(i) $S = 2$ can occur only when $x = 1$ and $y = 1$

So, $P(S = 2) = \frac{1}{36}$

(ii) $S = 3$ can occur only when; $(x = 1, y = 2)$ or $(x = 2, y = 1)$.

So, $P(S = 3) = \frac{2}{36}$

(iii) $S = 4$ can occur when; $(x = 1, y = 3)$, $(x = 2, y = 2)$, or $(x = 3, y = 1)$

So, $P(S = 4) = \frac{3}{36}$

$\therefore \quad P(S \leq 4) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{6}{36} = \frac{1}{6}$

Example 2.12 : A three member committee is constituted by selecting three persons at random from a group of 10 men and 8 women. Find the probability that both men and women are included in the committee.

Solution : The required event E can occur in two exclusive ways viz. (i) E_1, selection of 1 man and 2 women (ii) E_2, selection of 2 men and 1 woman.

$\therefore$ Probability of E = $\frac{640}{^{18}C_3} = \frac{640 \times 2 \times 3}{18 \times 17 \times 16} = \frac{40}{51}$

Example 2.13 : Three numbers are randomly selected from the set (1, 2, 3, 4, ..., 11, 12, 13}. Find the probability that the sum of the numbers is odd.

Solution : The desired event E can occur in the following exclusive ways-

(i) E_1, when all the three numbers is odd and other two are even.

(ii) E_2, when one of the numbers is odd and other two are even.

Number of ways of selecting 3 numbers from $13 = {}^{13}C_3 = 286$

Number of ways of selecting 3 odd numbers from the 7 odd numbers

of the set = ${}^{7}C_3 = 35$

Number of ways of selecting one odd number from the 7 and two even numbers from the 6 in the set = $7 \times {}^{6}C_2 = 7 \times 15 = 105$

So, required probability = $\frac{35 + 105}{286} = \frac{70}{143}$

Example 2.14 : Four balls are randomly selected from an urn containing 6 white, 4 black, and 5 red balls. What is the probability that balls of all colours are included in the choice ?

Solution : The desired event can occur in the following exclusive ways :-

$$\frac{C_2 \times 4 \times 5}{^{15}C_4}+\frac{6\times C_2 \times 5}{^{15}C_4}+\frac{6\times 4\times C_2}{^{15}C_4}$$

$$=\frac{15\times 5\times 4+6\times 6\times 5\times 6\times 4\times 10}{^{15}C_4}$$

$$=\frac{2\times 3\times 4}{15\times 14\times 13\times 12}\left[300+180+240\right]=\frac{48}{91}$$

Example 2.15 : Find the probability that a number randomly drawn from the set {1, 2, 3, 4, ..., 98, 99, 100} is divisible by 5 or 7.

Solution : Let A be the event that the number is divisible by 5 and B the event that the number is divisible by 7. Since there are 20 multiples of 5 in the set, $P(A) = \frac{20}{100}$.

Again, since there are 14 multiples of 7 in the set, $P(B) = \frac{14}{100}$. A number will be divisible by both 5 and 7 if it is divisible by 7 × 5 = 35 and there are two such number viz. 35 and 70. in the set. So, $P(AB) = \frac{2}{100}$.

$\therefore$ Requried probability = $P(A \cup B)$

$= P(A) + P(B) - P(A\,B)$

$$=\frac{20}{100}+\frac{14}{100}-\frac{2}{100}=0.32$$

Example 2.16 : A die is tossed twice. Find the probability that at least one of the scores is divisible by 3.

and $2 \times 2 = 4$ of these are favourable to AB.

$$\therefore \qquad P(A\,B) = \frac{4}{36} = \frac{1}{9}$$

$$\therefore \qquad \text{Required Probability} = \frac{1}{3} + \frac{1}{3} - \frac{1}{9} = \frac{5}{9}$$

Example 2.17 : A card is drawn from a well-shuffled deck. Find the probability that it is a red card or a picture card.

Solution :

Let A be the event that the card is a red card and B the event that it is a picture card.

Since there are 26 red cards and $4 \times 3 = 12$ picture cards (Jack, Queen, and King) in the deck,

$P(A) = 26 / 52$, $P(B) = 12/52$

Again, since half of the picture cards are red, $P(AB) = 6/52$

So, required probability $= P(A \cup B) = P(A) + P(B) - P(AB)$

$$= \frac{26}{52} + \frac{12}{52} - \frac{6}{52} = \frac{8}{13}$$

Example 2.18 : If $P(A) = 0.5$, $P(B) = 0.9$, what are the possible limits of $P(AB)$?

Solution :

Since $AB \subseteq A$, $AB \subseteq B$, $P(AB) \leq P(A)$ and $P(AB) \leq P(B)$

Thus, $P(AB) \leq 0.5$ and $P(AB) \leq 0.9 \Rightarrow P(AB) \leq 0.5$

Again, $P(A \cup B) = P(A) + P(B) - P(AB)$

$$= 0.9 + 0.5 - P(AB) \leq 1$$

can shrink if some extra information is available about it. We explain this concept through two simple examples - (i) Suppose two balls are successively drawn at random from an urn containing two balls, one red (R) and one black (B). Here, the sample space is S = {RB, BR}, which has two elements and the order in which the balls are drawn is indicated in each. Suppose E is the event that the second ball is red. Then, P(E) = 1/2. Now suppose we already see that the first ball is red. Then the elements of the sample space for which the first ball is red is simply $S' = \{RB\}$, since the other element BR does not have the first ball red. Hence, the probability that the second ball is red given the first is red is zero; since the subset of elementary events in S' for which the second ball is also red is ϕ. Thus, the extra information about the colour of the first ball has changed the probability of E from 1/2 to '0' i.e. it is now known for sure that the event E cannot happen. (ii) Suppose two numbers are randomly selected from the set {1, 2, 3, 4, 5, 6, 7}. The sample space consists of $^7C_2 = 21$ equally likely cases. Let E be the event that the sum of the numbers is even. Number of cases favourable to E = $^4C_2 + {^3C_2} = 6 + 3 = 9$; since the sum can be even only when both the numbers are even or both the numbers are odd. This gives, P(E) = 9/21 = 3/7. Suppose, it is given that one of the selected numbers is odd. Under this condition, there are only 6 possibilities for the second number since one number is eliminated. Out of these 6, only the choice of an odd number from the remaining 3 odd numbers is favourable to E. So, the probability of E in the changed sample space is 3/6 = 1/2.

In both these examples, the information given reduces the sample space. The probability of E, when this information is given or known, changes accordingly. This gives rise to the concept of conditional probability of an event A given that the event B

In the special case where S has n equally likely cases and m of these are favourable to B, P(B) = m/n. Out of these m cases, favourable to B, which are relevant when B has happened, if m' are favourable to A, then -

$$P(A/B) = \frac{m'}{m} = \left(\frac{m'}{n}\right) \div \left(\frac{m}{n}\right) = \frac{P(AB)}{P(B)}$$

as given above, Equation (2.2) can be rewritten as

$$P(AB) = P(B)\ P(A/B) \qquad (2.3)$$

Interchanging A and B, we get,

$$P(AB) = P(A)\ P(B/A) \qquad (2.3a)$$

Formulae (2.3) and (2.3a) are useful to calculate P(AB) if the computation of the relevant conditional probability is easier.

Results (2.2) and (2.3) are called the **MULTIPLICATION LAW OF PROBABILITY.**

To distinguish between the two, P(A) and P(A/B), both probabilities of A; are called unconditional and conditional probability of A respectively.

Independence - Events A and B are said to be independent, if P(A/B) = P(A). This means, a knowledge of B does not alter the probability of A. Substitution in (2.3) gives, if A and B are independent,

$$P(A\ B) = P(A)\ P(B) \qquad (2.4)$$

The above relation is often used as the definition of independence. The concept of independence of events can also be extended to more than two events. Thus three

If only (2.5), (2.6) and (2.7) are satisfied, the three events are said to be pairwise independent.

∴ None of the four conditions given in (2.5) to (2.8) is implied by the other conditions. Three events may be pairwise independent but not mutually independent. We illustrate this point with an example.

Example 2.18 : An urn contains four chits which bear the 3 - digit 0-1 sequences as follows :

Chit I - 000	Chit II - 011
Chit III - 101	Chit IV - 110

One chit is selected at random. Let A be the event that the first digit of the selected chit is 1, B the event that the second digit of the selected chit is 1, and C the event that the third digit of the selected chit is 1. Since the chits are equally likely to be selected,

$$P(A) = P(B) = P(C) = = \frac{2}{4} = \frac{1}{2}$$

$$\text{Again } P(AB) = P(AC) = P(BC) = \frac{1}{4}$$

This shows that conditions is (2.5), (2.6) and (2.7) are satisfied in this case.

But as there is no chit with all three digits equal to 1,

$$\therefore \quad P(ABC) = 0 \neq P(A)\,P(B)\,P(C) = \frac{1}{8}.$$

The concept of mutual independence of events can be extended to more than three events. For example, four events A_1, A_2, A_3, A_4 have to satisfy all the following conditions to be mutually independent.

(iii) Finally for all the 4 events [illegible]

Thus, (6+4 +1) = 11 conditions on the associated probabilities have to be satisfied for the mutual independence of 4 events. This implies that a knowledge of $P(A_1)$, $P(A_2)$, $P(A_3)$ and $P(A_4)$ will automatically determine the probabilities of the simultaneous occurence of any subset of these events by the conditions of independence.

Generalisation of the Multiplication Law of Probability

The multiplication law of probability, $P(AB) = P(A)P(B/A)$, can be extended to the cases of three or more events. For three events A, B, and C;

$$P(ABC) = P(AB)\ P(C/AB) = P(A)\ P(B/A)\ P(C/AB) \qquad (2.9)$$

The two steps of the above equation follow from the multiplication law for two events if we note that AB can be viewed as a single event. A direct extension of the above result follows, which can be written as -

$$P(A_1A_2A_3....A_n) = P(A_1)P(A_2/A_1)P(A_3/A_1A_2)....P(A_n/A_1A_2...A_{n-1}) \text{ .... } (2.10)$$

Where $P(A_i/A_1A_2...A_{i-1})$ means the conditional probability of A_i given the A_1, A_2, ...A_{i-1} have already happened. Formulae (2.9) and (2.10) are particularly useful to compute the simultaneous probability of three or more events which occur in a sequence, so that the conditional probabilities are easily computed.

Remarks :

(1) If A and B are independent events with $P(A) > 0$, $P(B) > 0$; then A and B cannot be disjoint.

Proof : Since $P(AB) = P(A)\ P(B) > 0$; $AB \neq \phi$.

(2) If A and B are disjoint events with $P(A) > 0$, $P(B) > 0$; then A and B cannot be independent.

Proof : Since AB and $A^c B$ are disjoint and $AB \cup A^c B = B$; we have :-

$$P(B) = P(AB) + P(A^c B)$$

$$= P(A)\,P(B) + P(A^c B) \qquad \text{(by independence)}$$

$$\therefore P(A^c B) = P(B) - P(A)\,P(B)$$

$$= P(B)\,[1 - P(A)] = P(B)\,P(A^c)$$

Thus A^c and B are also independent events.

Remark : If we repeat the above assertion to B and A^c i.e. A^c and B, we get :- If A and B are independent then A^c and B^c are also independent.

Example 2.20 : If A, B, C are mutually independent events; prove that (i) $(A \cap B)$ and C are independent (ii) $(A \cup B)$ and C are independent events.

Proof : (i) $P[(A \cap B) \cap C] = P(A \cap B \cap C)$

$$= P(A)\,P(B)\,P(C) \qquad \text{(by independence)}$$

$$= P(AB)\,P(C) \qquad \text{(by independence)}$$

which proves (i)

(ii) Similarly, $P[(A \cup B) \cap C]$

$$= P[(A \cap C) \cup P(B \cap C)] \qquad \text{(Distributive law)}$$

$$= P(A \cap C) + P(B \cap C) - P(A \cap B \cap C) \qquad \text{(Addition law)}$$

$$= P(A)\,P(C) + P(B)\,P(C) - P(A)\,P(B)\,P(C) \qquad \text{(Independence)}$$

$$= [P(A) + P(B) - P(AB)]P(C)$$

$$= P(A \cup B)\,P(C) \qquad \text{(Addition Law)}$$

which prooves (ii)

respectively. Thus, 2 R stands for (i) selection of Urn 2, and (ii) selection of a red ball from the selected Urn 2.

So, required probability = P(1R) + P(2R) $\quad$ (∵ 1R and 2 R are disjoint)

$$= P(1)\, P(R/1) + P(2)\, P(R/2)$$

$$= \frac{1}{2}\times\frac{2}{5}+\frac{1}{2}\times\frac{5}{7}=\frac{39}{70}$$

Here, P(1) represents the probability of selection of Urn 1, and P(R/1) represents the conditional probability of drawing a red ball when Urn 1 has been selected.

Example 2.22 : A problem is separately assigned to Ram, Sam and Hari for solution. The probabilities that Ram, Sam or Hari will solve the problem are 1/2, 2/5 and 1/3 respectively. What is the probability that the problem will be solved ?

Solution : Let A, B and C stand for the events that the problem will be solved by Ram, Sam and Hari respectively. These events do not influence each other since the problem is separately assigned to them. Hence they are independent events.

∴ Probability that the problem will not be solved by any one of them is

$$=P(A^cB^cC^c)=P(A^c)P(B^c)P(C^c)$$

$$=\left(1-\frac{1}{2}\right)\left(1-\frac{2}{5}\right)\left(1-\frac{1}{3}\right)=\frac{1}{5}$$

∴ Probability that the problem will be solved by at least one of them

$$= P(A\cup B\cup C)=1-P(A\cup B\cup C)^c$$

$$= 1-P(A^c\cap B^c\cap C^c) \qquad \text{(De Morgan's Law)}$$

$$= 1-\frac{1}{5}=\frac{4}{5}$$

$= P(AB^c C^c \cup A^c B C^c \cup A^c B^c C)$

$= P(AB^c C^c) + P(A^c B C^c) + P(A^c B^c C)$

$= P(A)P(B^c)P(C^c) + P(A^c)P(B)P(C^c) + P(A^c)P(B^c)P(C)$

$= 3p(1-p)^2$

Example 2.24 : Fi/e friends A, B, C, D and E buy the same kind of watch on the same day. The probability that each watch will develop a defect within 6 years is p. What is the probability that A's watch will develop a defect within 6 years and that his watch will be the first watch to develop a defect ?

Solution : The probability that a watch will not develop any defect in 6 years = 1 – p

∴ The probability that none of the 5 watches will develop any defect in 6 years = $(1-p)^5$, since the events are independent.

∴ The probability that at least one watch will develop a defect in 6 years = $1-(1-p)^5$,

Since the probability of developing the first defect is the same for each watch, the probability that it will be the watch of A = $\frac{1}{5}\left[(1-(1-p)^5\right]$, since there are five watches.

Required probability = $\frac{1}{5}\left[1-(1-p)^5\right]$

EXERCISE - 2(C)

1. Box 1 contains 5 chits bearing the letters A, B, C, D, E written in capital letters. Box 2 contains 4 chits bearing the letters a, b, c, d written in small letters. Two chits are drawn, one from each box, at random. Write down the sample space.
2. A die and a coin are tossed together. Write down the sample space.
3. Write down the sample space for four successive tosses of a coin.

6. Let the events (i), (ii) and (iii) of the above problem be denoted as A, B, and C. Write down all the elements of (i) $A-B$ (ii) $(B-C)\cup(C-B)$ (iii) $\overline{A}\cap\overline{B}\cap\overline{C}$

7. If n(A) denotes the number of elements in any set A;
Prove that :- (i) $n(A-B)=n(A\cup B)-n(B)$
(ii) $n(A)+n(B)=n(A\cup B)+n(A\cap B)$

8. Write down $E=(A\cup B\cup C\cup D)$ as the union of four disjoint events.

9. A coin is tossed three times. Write down all the possible outcomes and their associated probabilities.

10. A coin is tossed untill the first head is observed or untill three tosses are completed, whichever occurs earlier. Write down the sample space with the associated probabilities.

11. Three balls are successively drawn from a bag containing a white and b red balls. Write down the sample space and the associated probabilities ($a, b \geq 3$).

12. Two letters are randomly selected from the 26 letters of theEnglish alphabet. Find the probability that one vowel and one consonant will be included in the choice.

13. Two cards are randomly drawn from a full deck. Find the probability that at least one spade will be included in the choice.

14. Two numbers are randomly drawn from the set $\{1, 2, 3, \ldots, 2n-2, 2n-1, 2n\}$. Find the probability that their sum is an even number.

15. Three numbers are randomly selected from the set $\{1,2,3,\ldots,14,15,16\}$. What is the probability that their sum will be divisible by 3 ?

16. Find the probability that at least 3 out of 4 persons selected at random will be born on the same day of the week.

ace is drawn. Compute all relevant probabilities and verify :-

i. If events A, B and C are pairwise independent.

ii. If events A, B and C are mutually independent.

20. Three bags contain 3 red, 1 black; 2 red, 3 black; and 1 red, 3 black balls. One bag is selected at random and a ball drawn from it. Find the probability that it will be a red ball.

21. A coin is tossed. If a head is seen, then two dice are tossed but if a tail is seen, only one die is tossed. Find the probability that exactly one 6 score will be observed..

22. A die is tossed until a 6 score is observed. Find the probability that more than 5 tosses will be required.

23. A and B alternately throw a coin until a head is seen, A starting the process. Find the probability that A will be the first person to throw the first head. Also find the probability that the process will end with fewer than 8 throws.

24. If P(A) = 0.5, P(B) = 0.7 and P(AB) = 0.3; find $P(\overline{A}\ \overline{B})$.

25. A,B,C are independent events with probabilities P_1, P_2, P_3 respectively. Find the probability that (i) exactly two of these events will happen (ii) fewer than two of these events will happen.

26. A problem is given to Anil, Ram and Sam in that order for solution. The process is stopped as soon as a solution is obtained. The probabilities that Anil, Ram and Sam will solve the problem are 0.6, 0.5 and 0.3 respectively. Find the probability that the problem will be solved. Examine how this probability changes if the problem is given to all of them independently and simultaneously.

27. Three letters are randomly put into their addressed envelopes. Find the probability that at least one of them is sent to the correct addressee.

back his umbrella.

Hint : Use the method of the previous problem to find the probability of the complementary event.

2.12 SOME FURTHER TOPICS ON PROBABILITY

Inverse Probability

In our previous discussions, the conditional event A/B was defined as the occurrence of event A from among the elementary events which are favourable to B. While computing the probability P(A/B), the occurence of B was imposed as a pre-condition. This has a natural meaning if the event A follows the event B. The multiplication law of probability is useful to compute the probability of simultaneous occurrence of several events by placing them in a chronological order and computing the conditional probability of the next event at each stage.

But Bayes used the same formula to compute the probability of a preceding event when the event following it is observed. This may be useful if the record of what happened in the preceding stage is some how lost. We explain below, the theorem of Bayes, used to find such probabilities.

Bayes theorem - (i) Suppose an event B has been observed i.e. it has actually happened. (ii) Further suppose, $A_1, A_2, ..., A_k$ are mutually exclusive and exhaustive events which may precede B. That is, the event B has followed one of the A_i events (iii) Suppose the probabilities of $A_1, A_2, ..., A_k$ are denoted as $P(A_1), P(A_2), ...$ $P(A_k)$ respectively. Let the conditional probability of B, given that A_i has happened be denoted by $P(B/A_i)$. Then, Bayes theorem states that -

$(\because A_1, A_2, ...A_k$ are mutually exclusive and exhaustive)

$$= \frac{P(A_iB)}{P(\bigcup_{i=1}^{k} A_iB)}$$

$$= \frac{P(A_iB)}{\sum_{i=1}^{k} P(A_iB)}$$

$$= \frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{k} P(A_i)P(B/A_i)} \qquad \text{(by multiplication law)}$$

Remark :

The probabilities $P(A_1), P(A_2), ...P(A_k)$ are called prior probabilities of these events. These probabilities are modified as $P(A_1/B), P(A_2/B), ...P(A_k/B)$ after B has been observed. The extra information that B has happened modifies the 'prior' or 'a priori' probabilities. The modified probabilities $P(A_i / B)$, i= 1, 2, ... k are called 'posterior' or 'a posteriori' probabilities. Having observed B, $P(A_i)$ loses its significance and is substituted by $P(A_i/B)$.

The concept of finding the probability of an event on the basis of an event that followed it did not find favour with many contemporary scientists, who believed, it contradicts the basic principle of scientific enquiry. The posterior probabilities were called inverse probabilities. After prolonged resistance from a group of scientists called the non-Baylsians, Bayes theorem is now being widely used to solve problems where its application is appropriate. We give some illustrative example of its application.

Example 2.25 : Urns I and II respectively contain 1 white, 3 red; and 3 white, 2 red balls One urn was selected at random and a ball drawn from it at random. If it is a white ball, what is the probability that Urn I was originally selected ?

Also $P(B/A_1)=\frac{1}{4}$ and $P(B/A_2)=\frac{3}{5}$

$\therefore$ Required probability = $P(A_1/B)$

$$=\frac{P(A_1B)}{P(B)} \qquad \text{(Multiplication Law)}$$

$$=\frac{P(A_1B)}{P(A_1B)+P(A_2B)} \quad (\because A_1B \text{ and } A_2B \text{ are mutually exclusive and exhaustive})$$

$$=\frac{P(A_1)P(B/A_1)}{P(A_1)P(B/A_1)+P(A_2)P(B/A_2)}$$

$$=\frac{\frac{1}{2}\times\frac{1}{4}}{\frac{1}{2}\times\frac{1}{4}+\frac{1}{2}\times\frac{3}{5}}=\frac{5}{17}$$

Remark: Note that the original probability $P(A_1)=\frac{1}{2}$ has been reduced to $P(A_1/B)=\frac{5}{17}$; since observing the white ball reduces the probability of Urn 1 with a small proportion of white balls, being selected.

Example 2.26 : Similar bolts are manufactured by 3 machines A, B, and C in a factory. Machine A produces 45% of the bolts and 2% of these are defective; Machine B prodeces 30% of the bolts and 3% of these are defective, Machine C produces 25% of the bolts and 5% of these are defective. The bolts are assembled in one lot. What is the probability that a defective bolt found in the lot was produced by Machine A ?

Hence, $P(B/A_1) = 0.02$, $P(B/A_2) = 0.03$, $P(B/A_3) = 0.05$

Required probability $P(A_1/B) = \dfrac{P(A_1B)}{P(B)}$

$$= \frac{P(A_1B)}{P(A_1B)+P(A_2B)+P(A_3B)} \qquad (\because A_1B \text{, and } A_2B, A_3B \text{ are exclusive})$$

$$= \frac{P(A_1)P(B/A_1)}{P(A_1)P(B/A_1)+P(A_2)P(B/A_2)+P(A_3)P(B/A_3)}$$

$$= \frac{0.45 \times 0.02}{0.45 \times 0.02 + 0.30 \times 0.03 + 0.25 \times 0.05} = \frac{18}{61}$$

Example 2.27 : Urn 1 contains 4 red, 2 black and Urn 2 contains 1 red, 1 black balls. A ball is randomly transferred from Urn 1 to Urn 2. Then a ball randomly selected from Urn 2 was found to be black. Find the probability that the transferred ball was also black

Solution :

Let A be the event that the transferred ball was black and $\overline{A}$ the event that it was red. Let B be the event of a black ball being selected from Urn 2.

In the above prolem, $P(A) = \dfrac{2}{6} = \dfrac{1}{3}, P(\overline{A}) = \dfrac{2}{3}$

Also $P(B/A) = \dfrac{2}{3}$, $P(B/\overline{A}) = \dfrac{1}{3}$

Required probability $= P(A/B) = \dfrac{P(AB)}{P(B)}$ (Multiplication Law)

$$= \frac{P(A)\ P(B/A)}{P(A)P(B/A)+P(\overline{A})P(B/\overline{A})}$$

unknown. Such computation is not computing an inverse probability. We illustrate with an example below

Example 2.28 : Two urns contain balls of 3 colours as follows :

Urn 1 - 2 white, 3 red, 4 black

Urn 2 -3 white, 1 red, 2 black

One ball is randomly transferred from Urn 1 to Urn 2. Then a ball randomly drawn from Urn 2 is seen to be white. What is the probability that the next ball drawn from Urn 2 will also be white ?

Solution : Let A_1, A_2, A_3 respectiviely denote the events of a white, red or black ball being drawn fron Urn 1. Let B be the event that a ball randomly drawn from Urn 2 is white. Let C be the event that the next ball drawn will be white, the probability of which is required.

$$\text{Now, } P(A_1)=\frac{2}{9},\ P(A_2)=\frac{3}{9},\ P(A_3)=\frac{4}{9}$$

$$\text{Also } P(B/A_1)=\frac{4}{7},\ P(B/A_2)=P(B/A_3)=\frac{3}{7}$$

$$P(C/A_1B)=\frac{3}{6},\ P(C/A_2B)=P(C/A_3B)=\frac{2}{6}$$

$$\therefore \text{ Required probability} = P(C/B) = \frac{P(BC)}{P(B)} \qquad \text{(Multiplication law)}$$

$$= \frac{P\left(\bigcup_{i=1}^{3} A_iBC\right)}{P\left(\bigcup_{i=1}^{3} A_iB\right)} \qquad (\because A_1, A_2 \text{ or } A_3 \text{ can happen in first draw})$$

$$= \frac{\frac{}{9}\times\frac{}{7}\times\frac{}{6}+\frac{}{9}\times\frac{}{7}\times\frac{}{6}+\frac{}{9}\times\frac{}{7}\times\frac{}{6}}{\frac{2}{9}\times\frac{4}{7}+\frac{3}{9}\times\frac{3}{7}+\frac{4}{9}\times\frac{3}{7}}$$

$$= \frac{24+18+24}{6(8+9+12)} = \frac{11}{29}.$$

2.13 MUTUAL INDEPENDENCE OF SEVERAL EVENTS

Definition : A set of events $A_1, A_2, \ldots, A_k$ are mutually independent if the simultaneous occurrence of any subset of r of these events does not depend upon the happening or not happening of each of the remaining (k - r) events (r = 1, 2, 3, ..., k). This means, the probability of occurrence of any subset of these events depends only on the events included in the subset. It does not change with the happening or not happening of other events outside this subset. For example, for four events A_1, A_2, A_3, A_4, mutual independence requires -

$$P(A_1/A_2\,A_3\,A_4) = P(A_1/\,A_2\,\overline{A}_3\,A_4) = P(A_1\,/\,\overline{A}_2\,\overline{A}_3\,A_4) = P(A_1\,/\,\overline{A}_2\,\overline{A}_3\,\overline{A}_4) \text{ etc.}$$

We can simply write the above value as $P(A_1)$.

Also, $P(A_1\,A_2\,/\,A_3\,A_4) = P(A_1\,A_2\,/\,\overline{A}_3\,A_4) = P(A_1\,A_2\,/\,A_3\,\overline{A}_4) = P(A_1\;A_2\,/\,\overline{A}_3\,\overline{A}_4)$.

The above commom value may be simply written as $P(A_1\,A_2)$.

This again may be obtained as $P(A_1)\,P(A_2)$. This is implied in the definition.

Thus, if $A_1, A_2, \ldots, A_n$ are mutually independent, then all conditional probabilities are equal to the corresponding unconditional probabilities. Also unconditional probabilities of the simultaneous occurrence of a subset of events is equal to the product of the probabilities of the individual events in the subset. Thus, if $A_1, A_2, A_3, \ldots, A_n$ are mutually independent, then -

$$P(A_{i1} A_{i2} \dots A_{in}) = P(A_{i1})\, P(A_{i2}) \dots P(A_{in}) \dots \qquad (2.\ 11)$$

Remark 1 : If $A_1, A_2, \dots, A_n$ are mutually independent events, then events $B_1, B_2, \dots B_n$ are also mutually independent, where each B_i is either A_i or $\overline{A}_i$ (A_i^c).

Remark 2 : If $A_1, A_2, \dots, A_n$ are mutually independent events, any subset of these events are also mutually independent.

The above two results are useful to compute the probability of an event which happens if some of the above events happen and some others do not happen.

Example 2.29 : A_1, A_2, A_3, A_4 are mutually independent events with probabilities $P_1 . P_2$, $P_3 . P_4$ respectively. Find the probability of the following events -

(i) Exactly two of these events happen.

(ii) At least three of these events happen.

(iii) At least one of the events A_1, A_2, A_3 happens.

Solution : (i) E, the event that exactly two of them happen can be written as the union of six mutually exclusive events -

$$E = A_1 A_2 \overline{A}_3 \overline{A}_4 \cup A_1 \overline{A}_2 A_3 \overline{A}_4 \cup A_1 \overline{A}_2 \overline{A}_3 A_4 \cup \overline{A}_1 A_2 A_3 \overline{A}_4 \cup \overline{A}_1 A_2 \overline{A}_3$$
$$A_4 \cup \overline{A}_1 \overline{A}_2 A_3 A_4$$

So, required probability P(E)

$$= P(A_1 A_2 \overline{A}_3 \overline{A}_4) + P(A_1 \overline{A}_2 A_3 \overline{A}_4) + P(A_1 \overline{A}_2 \overline{A}_3 A_4) + P(\overline{A}_1 A_2 A_3 \overline{A}_4) +$$
$$P(\overline{A}_1 A_2 \overline{A}_3 A_4) + P(\overline{A}_1 \overline{A}_2 A_3 A_4)$$
$$= p_1 p_2 (1-p_3)(1-p_4) + p_1 p_3 (1-p_2)(1-p_4) + p_1 p_4 (1-p_2)(1-p_3) + p_2 p_3$$
$$(1-p_1)(1-p_4) + p_2 p_4 (1-p_1)(1-p_3) + p_3 p_4 (1-p_1)(1-p_2)$$

$$= p_1 p_2 p_3 (1-p_4) + p_1 p_2 p_4 (1-p_3) + p_1 p_3 p_4 (1-p_2) + p_2 p_3 p_4 (1-p_1)$$
$$+ P_1 P_2 P_3 P_4$$

(iii) By mutual independence of A_1, A_2, A_3:

$$P(\overline{A}_1 \overline{A}_2 \overline{A}_3) = P(\overline{A}_1)(\overline{A}_2)(\overline{A}_3) = (1-p_1)(1-p_2)(1-p_3)$$

$\therefore$ Required probability $= P(A_1 \cup A_2 \cup A_3)$

$$= 1 - P(\overline{A}_1 \overline{A}_2 \overline{A}_3)$$
$$= 1 - (1-p_1)(1-p_2)(1-p_3).$$

Example 2.30 : n persons take one chance each to hit a shooting target. The probability that the ith. person hits the target is p_i, $i = 1, 2, ..., n$. Find the probability that -

(i) no one hits the traget (ii) someone hits the target.

Solution : Let A_i be the event that ith. person hits the target, $i = 1, 2, ..., n$.

Then $P(A_i) = p_i$. These events are obviously mutually independent. So,

(i) P(no one hits the traget) $= P(\overline{A}_1 \overline{A}_2 ... \overline{A}_n) = P(\overline{A}_1) P(\overline{A}_2) ... P(\overline{A}_n)$

$$= (1-p_1)(1-p_2)...(1-p_n)$$

(ii) P(some one hits the target) $= P(A_1 \cup A_2 .... \cup A_n)$

$$= 1 - P(\overline{A}_1 \overline{A}_2 ... \overline{A}_n) = 1 - P(\overline{A}_1) P(\overline{A}_2) ... P(\overline{A}_n)$$
$$= 1 - (1-p_1)(1-p_2)...(1-p_n)$$

meeting at 4pm., the probability that he will reach home by 8pm. is 0.7. If he is at home to-day by 8pm, what is the probability that he had a meeting at 4pm, to-day ?

Hint : Let A be the event that Raj has a meeting at 4pm, and B the event that he is at home by 8pm, Required probability is -

$$= P(A/B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B/A)}{P(A)P(B/A) + P(\overline{A})P(B/\overline{A})} = \frac{0.4 \times 0.2}{0.4 \times 0.2 + 0.6 \times 0.7} = 0.16$$

3. Urn 1 contains a white, b black balls and Urn 2 contains c white, d black balls. One urn is chosen at random and a ball drawn from it. If it is white, what is the probability that originally urn 1 was selected ?

4. Urn 1 contains 2 white, 4 black balls and both urns 2 and 3 contain 4 white, 1 black ball each. One urn is selected at random and a ball drawn from it. If it is black, find the probability that urn 1 was originally selected.

5. An urn contains 3 white and 5 black balls. A ball is selected at random and two balls of the selected colour put back in the urn. Then a second ball is drawn from the urn. If this is black, what is the probability that originally a white ball was selected ?

6. Urn 1 contains 2 red, 4 white and urn 2 contains 3 red, 2 white balls. One randomly selected ball is transferred from urn 1 to urn 2. Then a randomly selected ball from urn 2 was seen as white. What is the probability that a white ball was transferred to urn 2 ?

7. In the previous problem, what is the probability that the next ball drawn from urn 2 after a white ball has already been drawn, will also be white ?

8. A cutting machine cuts plywood 70% of the time it is used, and cuts log wood in the balance 30% of the time. The probability that the machine will need adjustment by a mechanic is 1/50 when cutting a plywood, and 1/20 when cutting a log wood. If a

each question; find the probability that for a particular question, which has been correctly answered, the candidate knew the correct answer.

10. A, B, C are independent events with probabilities p_1, p_2 and p_3 respectively. Find expressions for (i) $P(A \cup B \cup C)$ (ii) $P(AB \cup AC)$

11. $A_1, A_2, \ldots, A_n$ are independent events with probabilities $p_1, p_2, \ldots, p_n$ respectively. Find the probability that at least one of these events happens.

12. $A_1, A_2, \ldots, A_n$ are independent events with the same probability p of happening. Find the probability that (i) exactly two of these events happen (ii) More than $(n-2)$ of these events happen.

13. Four boxes contain 3 red, 3 white; 2 red, 3 white; 3 red 2 white; and 1 red, 4 white balls respectively. One ball is randomly selected from each urn. Find the probability that there will be 2 white and 2 red balls in the choice.

14. A, B, C, D are independent events such that P(A)= P(B) =P, P(C)=P(D)= (1 – P). Show that the probability that exactly two of them happen is -

$$[(1-2P+2P^2)^2+2P^2(1-P)^2]$$

15. In the random selection of a card from a full deck, events A, B, C, D are defined as follows :-

(i) A happens if the selected card is an ace.

(ii) B happens if the selected card is a spade.

(iii) C happens if the selected card is red.

(iv) D happens if the selected card is a picture card i.e. a king or a queen or a jack. Compute the probabilities of A, B, C, D : all pairs of events AB, AC, etc; all triplets

MISCELLANEOUS EXERCISE - 2 (I)

1. In tossing of a die, give examples of :-

   (a) Three mutually exclusive events

   (b) Two mutually exhaustive events

   (c) Two events which are not exclusive.

2. In tossing of two dice, give an example of three events which are mutually exhaustive, but no two of which are exclusive.

3. Two numbers are randomly selected from the set {1, 2, 3, ..., 8, 9, 10}. Give an example of three mutually exclusive and exhaustive events.

4. In tossing of a die, give example of 3 equally likely events A, B, C such that A, C and B, C are exclusive but A, B are not exclusive.

5. In tossing of two coins, give examples of three events that are mutually exhaustive; but no two of which are equally likely.

6. In selection of a number from a set {1, 2, 3, ..., 18, 19, 20}, give examples of four mutually exclusive, exhaustive, and equally likely events.

7. U represents the set of all integers n, such that $10 \leq n \leq 20$. Write down the following subsets of U (i) A, the set of all prime numbers (ii) B, the set of all odd numbers, (iii) C, the set of all multiples of 5.

8. In the above problem write down the sets (i) $A \cup B$ (ii) $B^c \cup C$ (iii) $(A-B)$ (iv) $(B-A)$

9. $U = \{x : 0 \leq x \leq 1\}$. Give example of three disjoint sets A, B, C such that $A \cup B \cup C = U$

(i) Exactly one of the three events happens.

(ii) At least two of the three events happen.

(iii) All or none of the three events happen.

12. If A, B, C are events, write down the following events in descriptive language :-

(i) $(A \cup B \cup C)^c$ (ii) $\overline{A}BC \cup A\overline{B}C \cup AB\overline{C}$ (iii) $(A \cup B) - C$

13. Three letters are randomly selected from the set {a, b, c, d, e}. Write down the sample space with the associated probabilities.

14. A letter and a number are randomly selected from the sets {A, B, C, D} and {1, 2, 3} respectively. Write down the sample space with the associated probabilities.

15. A coin is tossed twice. Find the probability that both head and tail are observed.

16. A die is tossed twice. Find the probability that the sum of the scores exceeds 10.

17. Two persons are selected at random from a group of 6 men and 5 women. Find the probability that both of them are of the same sex.

18. Two numbers are randomly selected from the set {1, 2, 3, ..., 10, 11}. Find the probability that their sum is even.

19. A coin is tossed 4 times. Find the probability that the number of heads and tails will be equal.

20. Three numbers are randomly selected from the set {1, 2, 3, ..., 18, 19, 20}. Find the probability that the sum of the numbers when divided by 3 leaves a remainder 1.

21. Urn 1 contains 2 white, 4red balls and Urn 2 contains 5 white, 1 red balls. A coin is tossed. If it falls with head side up, a ball is randomly selected from urn 1, otherwise a ball is randomly selected from urn 2. Find the probability that a white ball wil be selected.

25. A number is randomly selected from the set {1, 2, 3, ..., 98, 99, 100}. Find the probability that it is divisible by 4 or 7. (+2, 2006)

26. In a residential area of Bhubaneswar with 200 households, 110 households buy an English newspaper and 70 buy a newspaper in a regional language. 30 of the households buy both the newspapers. What is the probability that a household selected at random does not buy any newspaper ?

27. Prove that for any events A, B, C :-

$$P(A \cup B \cup C) \geq P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC)$$

28. A bag contains 4 white, 3 red, and 2 black balls. Two balls are successively drawn from the bag at random. If the first ball is white, find the probability that the second ball is also white.

29. Two bag contain 3 red, 4 white and 4 red, 6 white balls. One ball is drawn from each bag. What is the probability that exactly one white ball will be included in the choice ?

30. Two urns 1 and 2 contain 1 red, 1 black and 2 red, 4 black balls respectevely. A coin is tossed. If a head is seen, a ball is drawn from Urn 1; if a tail is seen, a ball is drawn from Urn 2. What is the probability that a red ball will be drawn ?

31. If $P(A) = P(B) = 0.7$, $P(A \cup B) = 0.9$; find $P(B/A)$

32. If $P(A) = 0.8$, $P(B) = (B/A) = 0.3$; find $P(A \cup B)$

33. A and B are independent events with $P(A) = P(B) = p$.

Obtain an expression for $P(A \cup B)$

34. A, B, C are pairwise independent events with $P(A) = P(B) = P(C) = p$.

If $A \cap B \cap C = \phi$; find $P(A \cup B \cup C)$. Also find its maximum value when $0 \leq p \leq 1$.

37. An association has 14 men and 11 women members. One member was selected at random to attend a conference. Meanwhile, a request was received for a second name and a second person was selected at random. Find the probability that the second selected member is a woman.

38. From an urn containing 1 red, 2 white balls; a ball is transferred at random to urn 2 containing 'a' white, 'b' red balls. Then a ball is selected from urn 2. Find the probability that it is red.

39. A, B, C are independent events, each with probability p. Find an expression for $[P(A \cup B \cup C) - P(A \cup B)]$.

40. n mutually independent events $A_1, A_2, \dots, A_n$ have probabilities $p_1, p_2, \dots, p_n$ respectively. Find the probability that exactly one of them happens.

41. Three balls are randomly selected from a box containing 4 red and 5 white balls. Find the probability that a ball selected at random from the 3 selected balls is white.

42. A, B, C are mutually independent events with P(A) = 0.6, P(B) = 0.5, and P(C) = 0.8. Find $P(A \cup B \cup C)$ (+2, 2007)

43. $A_1, A_2, \dots, A_n$ are pairwise independent events and the probability of each event is p. If no three of these events can happen together, obtain an expression for $P\left(\bigcup_{i=1}^{n} A_i\right)$.

44. Out of two similar envelopes, the first contains 1 red, 7 white; and the second 3 red, 3 white chits. One chit is randomly transferred from the first envelope to the second. Then a chit drawn randomly from the second envelope was white. Find the probability that the colour of the transferred chit was white.

bolt is drawn at random from the product and found to be defective. What is the probability that it was manufactured by machine B ? (+2, 2006)

2.14 RANDOM VARIABLE : MATHEMATICAL EXPECTATION :

In the previous sections, probability was defined as a set function i.e. it was defined on subsets of a Universal set. Here, sets have a very general meaning of being a collection of any type of objects, the collection being well-defined. Our sets may be a group of students or plants, animals or furnitures. Let th.. elements of a finite set be denoted as $\{e_1, e_2, ..., e_n\}$

Random variable : A random variable X associates a number (real number) with every element of a set.

Let $x(e_i)$ be the number associated with e_i. Such association may be natural like the height of a tree in case of trees in a plantation, or mark secured by an individual student in case of a group of examinees, or number of children for a group of families. Thus, the random variable X transforms the element of the sample space $S = \{e_1, e_2, ..., e_n\}$ to n real numbers $\{X(e_1), X(e_2), ..., X(e_n)\}$. These n values need not be all different. Let the number of distinct values in this set of n values be denoted by $\{Z_1, Z_2, ..., Z_k\}$ where $k \leq n$. Each Z_i combines those values of $X(e_i)$ which are equal to Z_i. We can thus define a random variable X as a mapping from a set to the real line. We define a variable Z which takes the values $Z_1, Z_2, ..., Z_k$. Let $p(e_i)$ be the probability associated with e_i. We define P_i, the probability associated with Z_i as follows :-

$$P_i = P(Z = Z_i) = \sum p(e_i)$$

where the summation on the right extends over those e_i's for which $X(e_i) = Z_i$. Thus the random variable X transforms the sample Space $S=\{e_1, e_2, ..., e_n\}$ with associated where H and T represent the head side and tail side of the coin in the three successive tosses.

Let $X(e_i)$ denote the number of heads in the elementary event e_i, $i = 1, 2, 3, \ldots, 7, 8$. Then the $X(e_i)$ values obtained from the above sample space are -

$$\{3, 2, 2, 1, 2, 1, 1, 0\}$$

There are 4 distinct values in the above set which give -

$$(Z_1, Z_2, Z_3, Z_4) = \{3, 2, 1, 0\}$$

each Z_i having the corresponding value. For an unbiassed coin a probability $\frac{1}{8}$ can be assigned to each element of S.

So, $P(Z = Z_1) = P(Z = 3) = P(H\,H\,H) = \frac{1}{8}$

since 3 heads occur only in the element H H H.

$P(Z = Z_2) = (Z = 2) = P(H\,H\,T) + P(H\,T\,H) + P(T\,H\,H) = \frac{3}{8}$

since 2 heads occur in 3 elements H H T, H T H, and T H H.

Similarly, $P(Z = Z_3) = P(Z = 1) = \frac{3}{8}$

and $P(Z = Z_4) = P(Z = 0) = \frac{1}{8}$

The above table is called a probability distribution, where the total probability '1' is distributed among the values of Z.

(ii) In tossing of 2 dice, the sample space can be written as -

S = {(1,1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2,6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5,1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

where the pair of numbers represent the scores on the two dice in a definite order, say first ans second die. Let $X(\bar{e}_i)$ denote the total score of the two dice, which gives the following values of $X(e_i)$ from the elements of S.

{2, 3, 4, 5, 6, 7, 3, 4, 5, 6, 7, 8, 4, 5, 6, 7, 8, 9, 5, 6, 7, 8, 9, 10,
6, 7, 8, 9, 10, 11, 7, 8, 9, 10, 11, 12}

We note that there are 11 distinct values here, which gives -

$$\{Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8, Z_9, Z_{10}, Z_{11}\}$$
$$= \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

where each Zi has the corresponding value in the right hand set.

For two unbiassed dice a probability $\frac{1}{36}$ can be associated with each element of S.

So, $P(Z = Z_i) = P(Z = 2) = P[(1, 1)] = \frac{1}{36}$

since a total score of 2 occurs only for the pair (1, 1).

Similarly,

$$P(Z = Z_2) = P(Z = 3) = P[(1, 2) \cup (2, 1)] = \frac{1}{36} + \frac{1}{36} = \frac{2}{36}$$

$P(Z = 10) = \frac{3}{36}$, $P(Z = 11) = \frac{2}{36}$, $P(Z = 12) = \frac{1}{36}$

We represent the possible values of Z and the associated probabilities in the table given below :

no. of heads (Z) :	2	3	4	5	6	7	8	9	10	11	12
probability :	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

This is the probability distribution of the variable Z.

Remarks : (1) The above discussions also apply to situations where the number of elements in S is countable infinity i.e. there are infinitely many elements in S which can be arranged in a sequence. Here, X or Z can take a countable number of values. If Z takes a countable infinity number of values, then, a tabular representation of the infinite number of values is not possible. But a probability distribution can be described by a mathematical sequence. For example, if the number of tosses required to obtain the first head in a sequence of throws of a coin be Z, then $P(Z = i) = \frac{1}{2^i}$, $i = 1, 2, 3, \ldots\ldots \infty$.

(2) The set $\{X(e_1), X(e_2), \ldots, X(e_n)\}$ can be defined directly without any reference to the original events $\{e_1, e_2, \ldots, e_n\}$. Thus, the random variable X can be directly defined as $\{X_1, X_2, \ldots, X_n\}$ or rather as $\{Z_1, Z_2, \ldots, Z_k\}$ to prevent repeated values, where $X_i = X(e_i)$. If we are interested in Z, the original sample space may not have any relevance. Thus a random variable Z may be defined as one which takes values $\{Z_1, Z_2, \ldots, Z_k\}$ with associated probabilities $P(Z = Z_i) = P_i$, $i = 1, 2, 3, \ldots, k$, which $P_i \geq 0$ and $\sum_{i=1}^{k} P_i = 1$. This defines a probability distribution. We can extend the concept to a countable sequence $\{Z_1, Z_2, Z_3, \ldots\ldots\}$. Here, $(Z = Z_i)$, $i = 1, 2, \ldots$ are elementary events.

where $P_i \geq 0$, $\sum_{i=1}^{k} P_i = 1$

The definition can be extended to a countable number of values of Z.

ILLUSTRATIVE EXAMPLES : Given a probability distribution, probability of an event can be computed by adding the relevent probabilities. We give some examples.

Example 2.31 : The probability distribution of Z is given below :-

Z:	1	2	3	4	5	6	7	8	9	10
P_z	0.06	0.08	0.10	0.11	0.14	0.15	0.12	0.09	0.09	0.06

Find (i) $(Z \leq 4)$ (ii) $P(7 \leq Z \leq 10)$

(iii) find the smallest k, such that $P(Z \leq k) > 0.50$

Solution : (i) $(Z \leq 4) = P(Z = 1) + P(Z = 2) + P(Z = 3) + P(Z = 4)$

$= 0.06 + 0.08 + 0.10 + 0.11 = 0.35$

(ii) $P(7 \leq Z \leq 10) = P(Z = 7) + P(Z = 8) + P(Z = 9) + P(Z = 10)$

$= 0.12 + 0.09 + 0.09 + 0.06 = 0.36$

(iii) We cumulate the probabilities from below. i.e. from the smaller values :

This gives, $P(Z \leq 1) = 0.06$

$P(Z \leq 2) = 0.14$

$P(Z \leq 3) = 0.24$

$P(Z \leq 4) = 0.35$

$P(Z \leq 5) = 0.49$

$P(Z \leq 6) = 0.64$

So, k = 6

$$= \sum_{i=k}^{\infty} \frac{5^{i-1}}{6^{i}} = 5^{-1}\sum_{i=k}^{\infty}\left(\frac{5}{6}\right)^{i}$$

$$= 5^{-1}\sum_{j=0}^{\infty}\left(\frac{5}{6}\right)^{j+k} \qquad \text{where } j = (i-k)$$

$$= \frac{5^{k-1}}{6^{k}}\sum_{j=0}^{\infty}\left(\frac{5}{6}\right)^{j} = \frac{5^{k-1}}{6^{k}}\left(1-\frac{5}{6}\right)^{-1} = \left(\frac{5}{6}\right)^{k-1}$$

(ii) P(Z is an odd number) $\sum_{n=0}^{\infty} P(Z = 2n+1)$

$$= \sum_{n=0}^{\infty} \frac{5^{2n}}{6^{2n+1}} = \frac{1}{6}\sum_{n=0}^{\infty}\left(\frac{5}{6}\right)^{2n}$$

$$= \frac{1}{6}\sum_{n=0}^{\infty}\left(\frac{25}{36}\right)^{n} = \frac{1}{6}\left(1-\frac{25}{36}\right)^{-1} = \frac{6}{11}$$

2.15 CONTINUOUS RANDOM VARIABLES

In the preceding section, we have discussed about random variables which take a finite or countable number of values. Such variables are called "Discrete random variables". Discrete variables mostly take non-negative integral values which may extend upto infinity.

But there are many variables which can take any value in an interval of the real line. Height of a person, weight of an apple or a mobile phone, the distance by which an arrow misses a target are examples of such variables called "continuous random variables". They are comparable to continuous mathematical variables but have an intervals, so that questions of the above type may be answered.

Let a random variable X takes all values in an interval (A, B). Obviously, P(A<X<B)= 1. But probabilities are not located at individual points in this interval because there are uncountable number of points in the interval and their probabilities must add upto 1. So, only zero probability can be associated with points. It really does not make sense to talk about probability of an event that a person is exactly 175 cms tall because any microscopic deviation from 175 cms makes the outcome unfavourable to this event. As such, the probability of a person being exactly 175 cms tall is zero. It thus, makes sense to talk about the probability of X lying in a subinterval(a, b) of (A, B). We can ask, what is the probability that X lies in the interval (a, b) i.e. what is P(a <x<b) ?

Since the axiomatic approach, based on certain assumptions gives the best definition of probability; probabilities in intervals can be computed from an abstract probability structure on (A, B). Such a model probability structure can be generated by a function f(x), defined in (A, B), which satifies the following requirements :-

(i) $f(x) \geq 0$, (ii) $P(A<X<B) = \int_A^B f(x).\, dx = 1$

The probability associated with any sub-interval (a, b) of (A, B) is -

(iii) $P(a<x<b) = \int_a^b f(x)\, dx$

Any f(x) satisfying (i) and (ii) above generates a probability structure on subsets of (A,B) as defined by (iii). This is an abstract model of a continuous probability distribution; which does not solve a particular problem. But an appropriately chosen f(x) can represent an observed random situation. The interval (A, B) includes situations where (i) $A = -\infty$ (ii) $B = +\infty$, or (iii) $A = -\infty$ and $B = +\infty$, so that all situations are covered.

same for both and is a real variable.

Example 2.33 : The probability density function of a random variable X is given by -

$$f(x) = kx(1-x), \qquad 0<x<1$$

where k is a constant. (i) Find k (ii) Compute $P(X \leq 0.7)$.

Solution : $\int_0^1 f_x(x)\,dx = k\int_0^1 x(1-x)\,dx$

$$= k\left(\frac{x^2}{2}-\frac{x^3}{3}\right)_0^1 = \frac{k}{6} = 1 \qquad \text{(By the requirement of pdf)}$$

gives k = 6

(ii) $P(X \leq 0.7) = \int_{0.7}^{1} 6x(1-x)\,dx$

$$= \left(3x^2-2x^3\right)_{0.7}^{1}$$

$$= 1-\left[3x(0.7)^2-2x(0.7)^3\right] = 0.216$$

Example 2.34 : Robert collects his cup of tea from the snacks counter at any time between 1p.m. and 2p.m. If a note is left for him at the counter at 1.40p.m to-day, find the probability that he will get the note while collecting his cup of tea to-day.

Solution : Since the probability of Robert's arrival at the counter at any point of time between 1p.m. and 2p.m. is the same, the probability may be uniformly spread over the whole interval 1p.m. to 2p.m. We difine t = 0 at 1pm and t = 1 at 2p.m., so that t = k $(0 \leq k \leq 1)$ means the point of time k hours after 1p.m. The arrival time may be represented by a constant pdf. over $0 \leq t \leq 1$, as follows : -

i.e. if $t > \frac{40}{60} = \frac{2}{3}$.

$\therefore$ Required probability = $\int_{2/3}^{1} dt = 1 - \frac{2}{3} = \frac{1}{3}$.

Remark : In the above computations, X represents a random variable and x represents a real variable.

2.16 : MATHEMATICAL EXPECTATION :

We give below, the definition of mathematical expection or expectation or expected value of a random variable X (r.v.X).

(i) If the r.v.X is discrete and takes the values $x_1, x_2, x_3, ...$ with associated probabilities $p_1, p_2, p_3, ...$ respectively, then the expectation of X, E(X) is defined as

$$E(X) = \sum p_i x_i \tag{2.11}$$

provided the series on the right hand side absolutely converges in case X takes an inifinite number of values with positive probability i.e. provided.

$$\sum p_i |x_i| < \infty \tag{2.12}$$

In both (2.11) and (2.12), the summation extends over all the x_i-values. Absolute convergence of $\sum p_i x_i$ ensures that any rearrangement of terms does not alter the sum. (ii) If X is a continuous variable in the interval (A, B) with p.d.f(x), then E(X) is similarly defind as -

$$E(X) = \int_A^B x\, f(x)\, dx \tag{2.11a}$$

depending upon whether the variable is discrete or continuous, where x_i in (2.11) is replaced by $g(x_i)$ or x in (2.11a) is replaced by g(x). The condition of absolute convergence has to be ensured for E[g(X)] to be meaningful.

Thus, for discrete X,

$$E[g(X)] = \sum p_i\, g(x_i) \tag{2.13}$$

provided the righthand side absolutely converges.

For a continuous r.v.X,

$$E[g(X)] = \int_A^B g(x)\, f(x)\, dx \tag{2.13a}$$

provided the intergral on the right-hand side absolutely converges.

The following properties directly follow from the definitions and may be verified by the reader.

Property 1 : If g(x) = C, a constant ; then E g(x) = C

Property 2 : If g(x) = aX + b, where a, b are constants, and E(X) is finite, then

$$E(aX + b) = aE(X) + b$$

Remarks : (1) $E(X^r)$, where r is a positive integer, is called the rth moment of X about the origin, and usually denoted as μ'_r

$E(X) = \mu'_1$ and μ'_2, μ'_3, μ'_4 are important constants associated with the distribution of X.

(2) $$\mu_2 = E[X - E(X)]^2 \tag{2.14}$$

is called the variance of X, written as Var (X).

Var (X) in the form (2.14a) is convenient for applications.

(4) $\text{Var}(aX + b) = a^2\ \text{Var}(X)$ (2.15)

Proof : $\text{Var}(aX + b) = E[(aX + b) - E(ax + b)]^2$

$$= E[aX - aE(X)]^2$$

$$= a^2\ E[X - E(X)]^2 = a^2\ \text{Var}(X).$$

Illustrative Examples :

Example 2.35 : If X denotes the score in the single toss of a die, find E(X).

Solution : $P(X = i) = 1/6.$ $i = 1, 2, 3, 4, 5, 6.$

$$\therefore \quad E(X) = \sum_{i=1}^{6} iP(X = i)$$

$$= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{7}{2}$$

Example 2.36 : A selects a number at random from the set {1, 2, 3, ..., 8, 9, 10}. If he selects an even number he gets Rs 2, but if he selects a multiple of 5, he gets Rs 10. No payment is made for drawing the other numbers. Find the expectation of A's gain.

Solution : Let money received by A be Rs X.

$$P(X = 2) = P\text{ (An even number which is not a multiple of 5 is selected)} = \frac{4}{10}$$

$$P(X = 10) = P\text{ (A multiple of 5 is selected)} = \frac{2}{10}$$

$$\text{So, } E(X) = 2 \times \frac{4}{10} + 10 \times \frac{2}{10} = 2.8$$

Find (i) E(X)   (ii) $E(X^2)$ and Var (X)

Solution : $E(X) = -3\times 0.1+(-1)\times 0.3+1\times 0.3+5\times 0.1$

$$=-0.3-0.3+0.3+0.5=0.2$$

$$E(X^2) = 9\times 0.1+1\times 0.3+1\times 0.3+25\times 0.1$$

$$=0.9+0.3+0.3+2.5=4.0$$

$$\text{Var}(X) = E(X^2)-[E(X)]^2$$

$$=4.0-(0.2)^2=3.96$$

Example 2.38 : A coin is tossed until the first head appears. Find the expectation of the number of tosses.

Solution : If the first head appears in the nth. toss, then the number of tosses X = n.

$$P(X=n)=\frac{1}{2^n}, \qquad n=1,2,3,.....$$

since a tail has to appear in the first (n-1) tosses.

$$\text{So,}\quad E(X)=\sum_{n=1}^{\infty}\frac{n}{2^n}$$

$$=\frac{1}{2}\sum_{n=1}^{\infty}\frac{n}{2^{n-1}}$$

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

Example 2.39 : Starting with Ram, Ram and Sam alternately throw a die. He who throws a 5 or 6 first gets a prize of Rs 10, and the game ends with the award of the prize. Find the expectation of Ram's gain. (+2, 2007)

Ram wins the game if he throws a 5 or 6 for the first time in the (2n + 1)th throw where n = 0, 1, 2, 3, ...

$$\text{P[Game ends in the (2n + 1)th throw]} = \left(\frac{2}{3}\right)^{2n} \times \frac{1}{3} = \frac{1}{3} \times \left(\frac{4}{9}\right)^{n}$$

since scores other than 5 or 6 have to be thrown in the first 2n throws and then a 5 or 6 throw must follow.

$$\therefore \quad \text{P(Ram wins)} = \sum_{n=0}^{\infty} \frac{1}{3} \times \left(\frac{4}{9}\right)^{n} = \frac{1}{3}\left(1 - \frac{4}{9}\right)^{-1} = \frac{3}{5}$$

$$\therefore \quad \text{E(Ram's gain)} = \text{Rs}10 \times \frac{3}{5} = \text{Rs } 6.$$

2.17 BIVARIATE PROBABILITY DISTRIBUTION :

In our earlier discussions, we have considered a random variable X which takes real values with associated probabilities. This concept can be extended to a random variable taking a pair of values. This pairing may be natural to the phenomenon under study. We give some simple examples - (1) If two dice numbered 1 and 2 are tossed together, the outcome of each toss may be recorded by a pair of numbers (X, Y) where X is the score on die 1 and Y the score on die 2. The pair (X, Y) takes $6 \times 6 = 36$ possible pairs of values. For two symmetrical dice, a probability 1/36 may be associated with every pair. This probability distribution can be given in the form of the following two-way table.

3	36	36	36	36	36	36	6
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
Total	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

The entry against the row with heading X = i and the column with heading Y = j gives the probability P(X = i, Y = j) i, j = 1, 2, ..., 6. The last column gives the probability distribution of X and the last row the probability distribution of Y. The probabilities total to 1. Such a table is called a bivariate probability table.

(2) Suppose we have observed the variables (i) X, the number of rooms and (ii) Y, the number of persons in 100 houses of a locality. Let the possible values of X be X = 2, 3, 4, 5, 6 and possible values of Y = 1, 2, 3, 4, 5, 6, 7. If there are n_{ij} houses with X = X_i and Y = Y_j, we have $\sum\sum n_{ij}$ = 100, the total number of houses. If one household be selected at random then P(X = X_i, Y = Y_j) = n_{ij} / 100 = P_{ij}. A typical bivariate probability table may be observed as follows :

6	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.07
Total	0.08	0.12	0.18	0.21	0.17	0.13	0.11	1.00

Here also, the distributions of probabilities of X and Y have been shown as totals in the last column and last row respectively.

(3) Suppose a number X is randomly selected from the group of numbers (0, 1, 1, 2, 2). If X = 0, Y is the sum of two numbers randomly selected from (1, 1, 1). If X = 1, Y is the sum of two numbers randomly selected from (0, 1, 2) and if X = 2, Y is the similar sum of two numbers drawn from (1, 2, 3). The bivariate probability distribution of (X, Y) can be computed and verified to be as given below -

X \ Y	1	2	3	4	5	Total
0	0	$\frac{1}{5}$	0	0	0	$\frac{1}{5}$
1	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	0	0	$\frac{2}{5}$
2	0	0	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{5}$
Total	$\frac{2}{15}$	$\frac{5}{15}$	$\frac{4}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	1

The probability distributions of X and Y, computed as totals from the bivariate probability distribution; are called the marginal distributions of X. or Y. Each row (column) gives the probability values of Y (X), when X (Y) is held constant. If these probability values are computed from the bivariate probability distribution by adding the probabilities of cells which are favourable to the event. For example $P(X + Y \geq 11)$ can be computed in example (1) above by adding the probabilities of the cells, (X, Y) = (5, 6), (6, 5), (6, 6). Similarly in example (2) $P\left(\frac{2x}{3} < y < \frac{3x}{2}\right)$ can be computed by adding the probabilities of the cells (X, Y) = (2, 2), (3, 3), (3, 4), (4, 3), (4, 4), (4, 5), (5, 4), (5, 5), (5, 6), (5, 7), (6, 5), (6, 6), (6, 7) and (6, 8). (ii) In Example (1) the values of X do not influence the values of Y since they are scores of separate dice. Hence the events $(X = X_i)$ and $(Y = Y_i)$ are independent and we have.

$$P(X = x_i, Y = y_j) = P(X = x_i)\ P(Y = y_j)$$

The above formula holds good for all situations where X and Y are independent random variables. This can be used to compute the cell probabilities from the marginal probabilities of X and Y. But this is not true in Example (2) as more persons will require more rooms to accommodate themselves. Here, variables X and Y are mutually dependent and this dependence is reflected in the cell probabilities.

Generally speaking, a discrete bivariate probability distribution is defined on a set of pairs of points $\{(x_i, y_j) : (i,j) \in S\}$ where S is a finite or countable set of pairs of numbers. Probability at $(x_i, y_i) = P(X = x_i, Y = y_j) = p_{ij} \geq 0$,

such that $\sum\sum p_{ij} = 1$. Usually $S = I \times J$, where I and J are two sets of consecutive integers, which may be finite or countable.

Continuous case : As in the case of one variable (univariate), the variables X and Y in the bivariate case can be continuous. In this case, probabilities are not located at points (x, y), but in sets $\{(X, Y) : x < X \leq x + dx, y < Y \leq y + dy\}$, where dx and dy are intervals

$$\iint_{D} f(x, y)dxdy = 1$$

But in the present text, we will not deal with continuous bivariate probability distributions since their manipulation will require the knowledge of calculus at a higher level.

A simple example of a continuous random variable X is when X takes all values in the interval [0, 1] with equal probability. Here, the probab·'ty that X takes any particular value x_0 is zero. Probability is defined in intervals with $P(0 \leq x \leq 1) = 1$. For any sub-interval (a, b) of [0, 1], $P(a<x<b) = (b - a)$. By analogy, the simplest continuous bivariate distribution has the probability density function -

$$f(x, y) = [(b_1 - a_1)(b_2 - a_2)]^{-1} \; a_1 \leq x \leq b_1, a_2 < y \leq b_2$$

which is a rectangular domain in the XY - plane. This rectangle has area $[(b_1 - a_1)(b_2 - a_2)]$ and any part there of carries a proportionate probability. This is called the bivariate uniform probability denisity function, which has useful applications.

2.18 MATHEMATICAL EXPECTATION OF g(X, Y) :

We limit our discussion to a discrete bivariate probability distribution defined on a set of pairs denoted by $D = \{(x_i, y_j) : (i, j) \in S\}$, Let I denote the set of values of i and J the set of values of j occurring in S. We extend the set S to all pairs I × J, and assign zero probability to those pairs which do not occur in S. Thus, let

$$p_{ij} = \begin{cases} P(X = x_i, Y = y_j), & \text{if } (i, j) \in S \\ o & \text{if } (i, j) \notin S, (i, j) \in I \times J \end{cases}$$

We define,

$$p_{i\cdot} = P(X = x_i) = \sum_{j} p_{ij}, \quad p_{\cdot j} = P(Y = y_i) = \sum_{i} p_{ij}.$$

by rearrangement of the terms. If the number of element in S is finite, no such condition of absolute convergence is applicable.

ADDITION LAW OF EXPECTATION :

Theorem 2.18.1 - For a pair of random variables (X, Y)

$$E(X+Y) = E(X) + E(Y)$$

provided all the expectations exist.

Proof : - $E(X+Y) = \sum_i \sum_j (x_i + y_j)\, p_{ij}$

$$= \sum_i \sum_j x_i\, p_{ij} + \sum_i \sum_j y_i\, p_{ij}$$

$$= \sum_i x_i \left(\sum_j p_{ij} \right) + \sum_j y_i \left(\sum_i p_{ij} \right)$$

$$= \sum_i x_i\, p_{i\cdot} + \sum_j y_i\, p_{\cdot j}$$

$$= E(X) + E(Y)$$

Note that the rearrangement of terms at each step is permited by the condition of absolute convergence.

Corollary 2.18.1 - If $(X_1, X_2, \ldots, X_n)$ be n random variables, then

$$E\left(\sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} E(X_i)$$

provided all the expectations exist.

Proof : - The result can be proved by induction on n.

Theorem 2.18.1 states that the corollary holds for n = 2.

Let the corollary be true for (n – 1).

$= \sum_{i=1} E(X_i) + E(X_n)$ by induction hypothesis

$$= \sum_{i=1}^{n} E(X_i)$$

2.19. INDEPENDENCE OF RANDOM VARIABLES

Definition : Random variables X and Y are independent if the events $(X=x_i)$ and $(Y=y_j)$ are independent. i.e

$$P(X=x_i, Y=y_j) = P(X=x_i).\ P(Y=y_j)$$

for every pair (i, j). Symbolically, this can be written as -

$$p_{ij} = p_i p_j \qquad \text{for all pairs } (i,j)$$

Definition : Random variables $X_1, X_2, \ldots\ldots, X_n$ are independent if the events

$(X_1 = x_1), (X_2 = x_2), \ldots\ldots, (X_n = x_n)$ are mutually independent i.e.

$$P(X_1 = x_1, X_2 = x_2 \ldots\ldots, X_n = x_n) = P(X_1 = x_1)\, P(X_2 = x_2), \ldots\ldots P(X_n = x_n)$$

for every set of possible values $(x_1, x_2, \ldots\ldots, x_n)$ of $(X_1, X_2, \ldots\ldots, X_n)$.

Generally, two or more random variables which do not seem to be connected through any common link can be regarded as independent. The heights of two different persons or the weights of three fish in a catch can be regarded as independent random variables. But the weights of two brothers or two animals of the same strain are not independent random variables as they have a common genetic link. Similarly, to day's maximum temperature will influence the maximum temperature of to-morrow due to continuity of the season, and these variables are not mutually independent.

MULTIPLICATION LAW OF EXPECTATION :

Theorem 2.19.1 : If random variables X and Y are mutually independent, then

$E(X\ Y) = E(X)\ E(Y)$, provided all the expectations exist.

$= E(X)\,E(Y)$

Corollary 2.19.1 : If $X_1, X_2, ..., X_n$ be mutually independent random variables, then.

$$E(X_1, X_2, ... X_n) = E(X_1)\,E(X_2) .... (X_n)$$

provided all the expectations exist.

Proof : The previous theorem asserts that the above corollary is true when n = 2. We can prove the corollary for any n by induction. Let the corollary be true for (n – 1) random variables $X_1, X_2, ... X_{n-1}$ then,

$$E(X_1, X_2, ... X_n) = E[(X_1, X_2, ... X_{n-1})\,X_n]$$

$$= E(X_1, X_2, ... X_{n-1})\,E(X_n) \qquad \text{(by independence)}$$

$$= E(X_1)\,E(X_2) .... E(X_{n-1})E(X_n) \qquad \text{(by induction hypothesis)}$$

2.20 VARIANCE OF SUM OF RANDOM VARIABLES

Mutually independent random variables :

Theorem 2.20.1 : If X and Y are mutually independent, then

$$\text{Var}\,(X + Y) = \text{Var}\,(X) + \text{Var}\,(Y) \qquad (2.17)$$

Proof :

$$\text{Var}\,(X + Y) = E(X + Y)^2 - [E(X + Y)]^2$$

$$= E(X^2 + Y^2 + 2XY) - [E(X) + E(Y)]^2 \qquad \text{(Theorem 2.18.1)}$$

$$= E(X)^2 + E(Y)^2 + 2E(XY) - [E(X)]^2 - [E(Y)]^2 - 2E(X)E(Y)$$

(by Corollary 2.18.1)

$$= [E(X^2) - \{E(X)\}^2] + [E(Y^2) - \{E(Y)\}^2]$$

$$[\because \; E(XY) = E(X)\,E(Y)]$$

$$= \text{Var}\,(X) + \text{Var}\,(Y)$$

variable $X_1, X_2, \ldots\ldots, X_{n-1}$. Then,

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \text{Var}\left(\sum_{i=1}^{n-1} X_i + X_n\right)$$

$$= \text{Var}\left(\sum_{i=1}^{n} X_i\right) + \text{Var}(X_n) \qquad \text{(by independence)}$$

$$= \sum_{i=1}^{n-1} \text{Var}(X_i) + \text{Var}(X_n) \qquad \text{(by induction hypotheses)}$$

$$= \sum_{i=1}^{n} \text{Var}(X_i)$$

Dependent random variables : -

Definition : If random variables X and Y are not independent, then covariance of (X, Y), Cov (X, Y) is given by,

$$\text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] \ldots \qquad (2.19)$$

If X and Y are independent, the covariance is zero. It is seen that

$$\text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = E[XY - XE(Y) - Y\,E(X) + E(X)\,E(Y)]$$

$$= E(XY) - E(X)\,E(Y) \qquad (2.19a)$$

which is a convenient from for applications.

Theorem 2.20.2 : For a pair of variables (X, Y),

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$$

Proof : $\text{Var}(X + Y) = E[(X + Y) - E(X + Y)]^2 \qquad (2.20)$

$$= E(X + Y)^2 - [E(X + Y)]^2$$

$$= E(X^2 + Y^2 + 2XY) - [E(X) + E(Y)]^2$$

Proof : $$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = E\left(\sum_{i=1}^{n} X_i\right)^2 - \left[E\left(\sum_{i=1}^{n} X_i\right)\right]^2$$

$$= E\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i<j}^{n}\sum_{=1} X_iX_j\right] - \left[\sum_{i=1}^{n} E(X_i)\right]^2$$

$$= E\sum_{i=1}^{n}(X_i)^2 + 2E\sum_{i<j}^{n}\sum_{=1}(X_iX_j) - \left[\sum_{i=1}^{n}\{E(X_i)\}^2 + {}^{n}2\sum_{i<j}\sum_{=1} E(X_i)\ E(X_j)\right]$$

$$= \sum_{i=1}^{n}\left[E(X_i^2) - \{E(X_i)\}^2\right] + 2\sum_{i<j}^{n}\sum_{=1}\left[E(X_iX_j) - E(X_i)E(X_j)\right]$$

$$= \sum_{i=1}^{n}\text{Var}(X_i) + 2\sum_{i<j}\sum_{=1}^{n}\text{Cov}(X_iX_j)$$

Remark : We have seen that if X and Y are independent, then Cov (X, Y) = 0.

The converse is not true i.e. Cov (X, Y) = 0 does not imply the independence of X and Y. For example if X takes values -2, −1, 0, 1, 2 with equal probability and $Y = X^2$, then -

$$\begin{aligned} \text{Cov}(X, Y) &= E(XY) - E(X)\,E(X^2) \\ &= E(X^3) - E(X)\,E(X^2) \\ &= 0 \qquad\qquad [\because\ E(X) = E(X^3) = 0] \end{aligned}$$

ILLUSTRATIVE EXAMPLES :

Example 2.40 : Two chits are sucessively drawn from an urn containing 3 chits bearing numbers 1, 2, 3. If X be the number on Chit - 1 and Y the number of chit - 2 represent the probability distribution of (X, Y) in a biavariate table.

bivariate table.

X \ Y	1	2	3
1	0	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	0	$\frac{1}{6}$
3	$\frac{1}{6}$	$\frac{1}{6}$	0

Example 2.41 : 40 students of a class were given 2 questions each carrying 4 marks. (X, Y) represents the score of a student in questions 1 and 2 respectively. The relative frequencies of (X, Y) pairs are given in the following bivariate table. Obtain (i) the marginal distribution of Y (ii) conditional distribution of X given Y = 3

X \ Y	0	1	2	3	4
0	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{1}{40}$	0	0
1	$\frac{1}{40}$	$\frac{3}{40}$	$\frac{2}{40}$	$\frac{2}{40}$	0
2	0	$\frac{3}{40}$	$\frac{7}{40}$	$\frac{2}{40}$	$\frac{4}{40}$
3	0	$\frac{1}{40}$	$\frac{4}{40}$	$\frac{4}{40}$	$\frac{1}{40}$
4	0	0	$\frac{1}{40}$	$\frac{2}{40}$	0

0	0	0
1	$\frac{2}{40}$	$\frac{2}{10}$
2	$\frac{2}{40}$	$\frac{2}{10}$
3	$\frac{4}{40}$	$\frac{4}{10}$
4	$\frac{2}{40}$	$\frac{2}{10}$
Total	$\frac{10}{40}$	$\frac{10}{10} = 1$

The conditional relative frequencies of X given Y = 3 are shown in column with heading R(x). Their sum is $\left(\frac{10}{40}\right)$. The R(x) figures have been multiplied by 4 to give a total of '1' in column P(X), which gives the conditional distribution of X when Y= 3.

Example 2.42 : From the following bivariate probability distribution compute

(i) $P(X = Y)$ (ii) $P(|X - Y| > 2)$

X \ Y	1	2	3	4	5	6
1	0.03	0.04	0.02	0.01	0	0
2	0.01	0.04	0.08	0.02	0.02	0
3	0.02	0.02	0.11	0.24	0.07	0.04
4	0.01	0.01	0.02	0.04	0.04	0.02
5	0	0	0.02	0.02	0.03	0.02

$+ P(x = 4, y = 1) + \sum_{j=1} P(x = 5, y = j)) = 0.01 + 0.02 + 0.04 + 0.01 = 0.08.$

Example 2.43 : From the following bivariate probability distribution, compute

(i) E(XY) (ii) E(|X – Y|)

X \ Y	1	2	3
1	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{1}{20}$
2	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{2}{20}$
3	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{3}{20}$

Solution :

XY :	1	2	3	4	6	9
P(XY) :	$\frac{3}{20}$	$\frac{4}{20}$	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{3}{20}$

$$E(XY) = 1 \times P(XY = 1) + 2 \times P(XY = 2) + 3 \times P(XY = 3) + 4 \times P(XY = 4)$$

$$+ 6 \times P(XY = 6) + 9 \times P(XY = 9)$$

$$= 1 \times \frac{3}{20} + 2 \times \frac{4}{20} + 3 \times \frac{2}{20} + 4 \times \frac{4}{20} + 6 \times \frac{4}{20} + 9 \times \frac{3}{20}$$

$$= \frac{1}{20}(3 + 8 + 6 + 16 + 24 + 27) = \frac{84}{20} = 4.2$$

(ii)

|X - Y| :	0	1	2
P(|X - Y|) :	$\frac{10}{20}$	$\frac{8}{20}$	$\frac{2}{20}$

Solution : Let $X_1, X_2, \ldots\ldots, X_n$ be the scores of the n tosses.

Let $S = \sum X_i$

$$E(X_i) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}\text{., } i = 1, 2, \ldots, n.$$

$$\therefore\ E(S) = E(\sum X_i) = \sum E(X_i) = nE(X_i) = 7n/2$$

Example 2.45 : n balls are drawn from a box containing a white and b black balls. Find the expected number of white balls.

Solution : Let $X_i = \begin{cases} 1 & \text{if the ball is white} \\ 0 & \text{if the ball is black} \end{cases}$

Then, number of white balls, $S = \sum_{i=1}^{n} X_i$

$$E(X_i) = 1.\ \frac{a}{a+b} + 0.\ \frac{b}{a+b} = \frac{a}{a+b}$$

$$\therefore\ E(S) = nE(X_i) = \frac{na}{a+b}$$

Example 2.46 : Box 1 contains 3 chits bearing numbers 2, 4, 6 and box 2 contains 4 chits bearing numbers 3, 5, 7 and 9. One chit is randomly selected from each box. Find the expectation of the product of the scores obtained.

Solution : Let the scores obtained from boxes 1 and 2 be X and Y respectively and the product of the scores P = XY. Since X and Y are independent, E(P) = E(X) E(Y).

Now, $E(X)\ \frac{2+4+6}{3} = 4$, $E(Y)\ \frac{3+5+7+9}{4} = 6.$

So, $E(P) = 4 \times 6 = 24.$

independent. Thus,

$$X_i = \begin{cases} a & \text{with probability } p \\ b & \text{with probability } (1-p) \end{cases}$$

$$\therefore\ E(X_i) = ap + b(1-p) = [b + (a-b)p]$$

(i) Sum of the scores $S = X_1 + X_2, \dots\dots + X_n$.

$$\therefore\ E(S) = nE(X_i) = n[b + (a-b)p]$$

(ii) Product of the scores $P = X_1, X_2, \dots\dots, X_n$.

$$\therefore\ E(P) = [E(X_i)]^n = [b + (a-b)p]^n$$

Example 2.48 : n trials are made, each resulting in a success or a failure with probability p and (1 - p) respectively. Find the expected value and variance of the number of successes.

Solution : Let

$$X_i = \begin{cases} 1 \text{ if the ith. trial is a success} \\ 0 \text{ if the ith. trial is a failure.} \end{cases}$$

Obviously, the number of successes $S = \sum_{i=1}^{n} X_i$

The X_i - variables are independent and $E(X_i) = p$.
$\therefore\ E(S) = np$.

Again $E(X_i^2) = p$, $Var\ (X_i) = E(X_i^2) - [E(X_i)]^2 = p - p^2 = p(1-p)$

$$\therefore\ Var\ S = \sum_{i=1}^{n} Var\ (X_i) = np(1-p)$$

Example 2.49 : m chits are randomly selected from the N chits of a box bearing numbers 1, 2, ..., N. If S be the sum of the numbers obtained, find (i) E(S) (ii) Var (S).

(i) $\therefore E(S) = E\left(\sum_{i=1}^{m} X_i\right) = \sum_{i=1}^{m} E(X_i) = \frac{1}{2}m(N+1)$

(ii) Since (X_i, X_j) take any pair of values from $\{1, 2, ..., N\}$

$$E(X_iX_j) = \frac{1}{N(N-1)}\sum_{i}\sum_{\neq j} X_iX_j$$

$$= \frac{1}{N(N-1)}\left[(\sum X_i)^2 - \sum X_i^2\right]$$

$$= \frac{1}{N(N-1)}\left[(1+2+3+....+N)^2 - (1^2+2^2....+N^2)\right]$$

$$= \frac{1}{N(N-1)}\left[\frac{N^2(N+1)^2}{4} - \frac{N(N+1)(2N+1)}{6}\right]$$

$$= \frac{N+1}{12(N-1)}\left[3N(N+1) - 2(2N+1)\right]$$

$$= \frac{N+1}{12(N-1)}\left[3N^2 - N - 2\right] = \frac{(N+1)(3N+2)}{12}$$

$$\therefore \text{Cov}\,(X_iX_j) = E(X_iX_j) - E(X_i)\,E(X_j)$$

$$= \frac{(N+1)(3N+2)}{12} - \frac{(N+1)^2}{4} = -\frac{N+1}{12}.$$

Again, $\text{Var}(X_i) = E(X_i^2) - [E(X_i)]^2$

$$= \frac{1^2+2^2....+N^2}{N} - \frac{(N+1)^2}{4}$$

$$= \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} = \frac{N^2-1}{12}.$$

$$= \frac{\quad}{12} - \frac{\quad}{12} = \frac{\quad}{12}$$

Exercise - 2(E)

1. X is the sum of two numbers randomly drawn from the set {1, 2, 3, 4}. Write down the probability distribution of X.

2. 4 balls are randomly selected from an urn containing 5 white and 6 black balls. Write down the probability distribution of the number of white balls.

3. Three numbers are randomly selected from {1, 2, 3, 4, 5}. Write down the probability distribution of the largest number selected.

4. The probability mass function of X is given by $P(X = K) = \frac{A}{K!}, K = 1, 2, 3, \ldots,$ where A is a constant. (i) Find A (ii) Find the probability that X takes an odd value.

5. X and Y are random observations, taken with replacement, from the same set {1, 2, 3, 4, 5}. Find the probability mass function of $Z = (X - Y)$.

6. The probability density function of X is given by $f(x) = kx^2(1-x), 0 \le x \le 1$. (i) Find k (ii) Find the probabilities $P(X<0.2)$, $P(X>0.5)$.

7. The p.d.f of X is given by $f(x) = kx,\ 0 \le x \le 2$, where k is a constant. Find (i) k (ii) $P(0.5 \le x \le 1.5)$

8. Random variable X takes values 1, 2, 3 with probabilities inversely proportional to these values. Determine (i) E(X) (ii) $E(X^2 + 2)$.

9. Two balls are randomly selected from an urn containing 4 white and 5 red balls. Find the expected number of white balls selected.

X	:	1	2	3	4	5	6
P(X)	:	.0.5	.15	0.40	0.20	.05	.15

12. Deepa throws a die until she gets a score of 3 or 6 for the first time. If she has to pay one rupee for every throw which is not a 3 or 6, find the expectation of her loss.

13. Three numbers X, Y, Z are randomly selected from the set {1, 2, ..., n}

Find E(X + Y + Z).

14. The table below gives the probability distribution of (X, Y)

X \ Y	1	2	3	4
1	0.04	0.18	0.00	0.16
2	0.16	0.05	0.04	0.18
3	0.00	0.07	0.05	0.07

(i) Find the marginal distribution of Y

Obtain (ii) P(Y>X) (iii) P(Y = X).

15. In selecting two cards from a well-shuffled deck, which of the following options give a higher expected return ?

(A) Prize of Rs 500 for selecting a king and a queen of the same suit. (B) Prize of Rs 350 for selecting two aces.

16. Ram draws a card from a well shuffled deck. If he gets a spade, he gets a prize of Rs. 20 and draws a second card from the remaining 51. If he draws a second spade, he gets Rs. 30 more and draws a third card from the remaining 50. If the third card is a spade, he gets Rs 50 more. No other payment is made for other choices. Find the expectation of his gain.

17. Numbers X and Y are randomly selected from the sets {1, 3, 5, ..., 2n + 1} and $\{1^2, 2^{2,} 3^2, ..., n^2\}$. Find (i) E(XY) and (ii) E(X + Y).

20. m different numbers $y_1, y_2, ..., y_m$ are selected at random from the set of numbers $\{x_1, x_2, ..., x_N\}$. If $S = \sum_{i=1}^{m} y_i$, obtain expressions for E(S), and Var (S) in terms of $\bar{x} = \sum_{i=1}^{N} x_i / N$ and $s^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2 / N$.

Miscellaneous Exercise - 2(II)

1. The letters of the word PROBABILITY are arranged randomly in a sequence. What is the probability that the last letter of the sequence will be A or O or I ?

2. What is the probability that a new year will start with a Saturday or Sunday.

3. A number is selected at random from the set {1, 2, 3, ..., 18, 19, 20}. Find the probability that it will be divisible by 3 or 7.

4. The numbers of the set {1, 2, 3, ..., 10, 11, 12} are randomly divided into two halves. Find the probability that these will be three even and three odd numbers in each half.

5. Three numbers are randomly selected from the set {1, 2, 3, ..., 10, 11, 12, 13, 14}. Find the probability that at least one double digit number will be included in the choice.

6. Three balls are randomly selected from a bag containing 3 red, 4 black and 4 white balls. Find the probability that at least two of them will be of the same colour.

7. If U = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}; give examples of (i) 3 sets which are mutually exclusive but not exhaustive (ii) 3 sets which are exhaustive but not mutually exclusive (iii) 3 sets which are both mutually exclusive and exhaustive.

8. Write down all elements of the following sets : - (i) All positive integers ≤ 10 which are the products of two prime factors (ii) All positive integers which can be expressed as $\frac{m}{n}$ wher m and n are positive integers with (m+n) = 9. (iii) All numbers of the from 3^n not exceeding 100, where n is a positive integer.

6). Write down the sample space with the associated probabilities.

12. A number is randomly selected from the first 50 positive integers. Find the probability that it is divisible by 6 or 7 or both.

13. Four balls are randomly selected from a bag containing 2 blue and 9 red balls. Find the probability that at least one blue ball is included in the choice.

14. If P(A) = 0.3, P(B) = 0.4, P($A \cup B$) = 0.58; prove that

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

15. A ball is selected at random from an urn containing 5 white, 6 red balls. Two balls of the selected colour are put back in the urn and a second ball is drawn. Find the probability that it is red.

16. Hari and Ram select two cards successively from a well-shuffled deck. Find the probability that Hari selects a spade and Ram selects an ace.

17. Henry will visit drug stores A, B and C to buy a medicine. The probabilities that he will get the medicine in these stores are 0.7, 0.6 and 0.5 respectively. If he visits the stores in this order, find the probability that he will get the medicine.

18. If P(A) = 0.75, P(B) = 0.62; Find the limits between which

(i) P($A \cup B$) and (ii) P(AB) will lie.

19. Which one of the following is true and which one is false ?

(i) $P(A/B) \geq P(A) \Rightarrow P(B/A) \geq P(B)$

(ii) $P(A) = P(B) \Rightarrow A = B$

(iii) $P(A) = 0 \Rightarrow A = \phi$

(iv) $P(B/A) = P(B/\overline{A}) \Rightarrow$ A and B are independent.

20. Events A and B are independent. If P(C) = 0, prove that A, B and C are mutually independent.

60% of the apples come from A, 25% from B and 15% from C. 5% of the apples from A, 6% from B and 8% from C are second quality. Find the probability that a randomly selected second quality apple has come from source A.

24. Urn 1 has twice as may red balls as white balls, and urn 2 has 2 white and 1 red balls. One ball is randomly transferred from urn 1 to urn 2 and then a ball randomly chosen from urn 2. If it is a white ball, find the probability that the transferred ball was also white.

25. A pair of numbers (X, Y) is randomly drawn from tne set {1, 2, 3, 4, 5}. Find the probability mass function of -

(i) $Z = (X + Y)$ (ii) $Z = |X - Y|$

26. X and Y are randomly drawn numbers from the sets {1, 2, 3} and {2, 3, 4, 5} respectively. Find the probability distribution of $Z = X^2 + Y$.

27. Random variable X has the probability mass function :-

$P(X = i) = K / 3^i$, $i = 1, 2, 3, \ldots$. Find K and E(X).

28. If X and Y be a pair of random variables, find an expression for Var (AX + BY) in terms of the variances and covariance of X and Y. Hence examine the significance of,

$$\text{Var}(AX + BY) = \text{Var}(AX - BY)$$

29. Find the expected number of failures preceding the first success in an infinite series of independent trials with constant probability of success p in each trial.

30. n letters serially numbered as 1, 2, 3, ..n are put into n envolopes also numbered 1, 2, 3, ..., n in a random manner. If the ith. letter is put into the ith. envelope, we get a match (i = 1, 2, ..., n). Find the expected value and the variance of the number of such matches.

31. Obtain an expression for Cov (a X + bY, cX + dY) in terms of the variances and covariance of X and Y.

15. $\frac{21}{143}$ 16. $\frac{35}{128}$ (ii) $\frac{1}{2}$ 17. $\frac{{}^{13}C_8}{{}^{20}C_8}$ 18. $\frac{52}{625}$

Exercise 2 (B)

1(a) 2, 3, 5, 7, 11, 13, 17, 19 (b) 1, 4, 9, 16, 25, 36, 49

(c) 20, 25, 30, 35, 40 (d) 21, 28, 35, 42, 49

(e) 8, 12, 18, 20

2 (i) [0, 5, 1.0] (ii) $[0.5, 0.6) \cup (0.8, 1]$ (iii) (0.4, 0.5)

3. (i) {19, 9, 4, 3, 1} (ii) {19, 9, 4, 3, 1.5, 1, 0.25, 0}

(iii) $\left\{\frac{19}{1}, \frac{18}{2}, \frac{17}{3}, \frac{16}{4}, \frac{15}{5}, \frac{14}{6}, \frac{13}{7}\right\}$

4. (i) {e} (ii) ϕ (iii) {i, j} (iv) {a, j}

Exercise 2 (C)

4. {5, 7, 9}, {3, 7, 9}, {3, 5, 9}, {3, 5, 7}, {1, 7, 9}, {1, 5, 9}, {1, 5, 7}, {1, 3, 9}, {1, 3, 7}, {1, 3, 5}

5 (i) {1, 3, 5, 7, 9, 11, 13, 15} (ii) {7, 8, 9, 10, 11, 12}

(iii) {3, 5, 6, 9, 10, 12, 15} (iv) {2, 4, 8}

6. (i) {1, 3, 5, 13, 15} (ii) {3, 5, 6, 7, 8, 11, 15} (iii) {2, 4, 14}

8. $A \cup (A^cB) \cup (A^cB^cC) \cup (A^cB^cC^cD)$

18. $p_1p_2, p_1+p_2-p_1p_2$ 19. $P(A)=\frac{1}{4}, P(B)=\frac{3}{13}, P(C)=\frac{1}{13}, P(AB)=\frac{3}{52}$,

$P(AC)=\frac{1}{52}$, $P(BC)=P(ABC)=0$. Only A and B are independent.

20. $\frac{7}{15}$ 21. $\frac{2}{9}$ 22. $1-\left(\frac{5}{6}\right)^5$ 23. (i) $\frac{2}{3}$ (ii) $\frac{255}{256}$ 24.0.10

25. (i) $p_1p_2(1-p_3)+p_1p_3(1-p_2)+p_2p_3(1-p_1)$

(ii) $(1-p_1)(1-p_2)(1-p_3)\left[1+\sum_{i=1}^{3}\frac{p_i}{1-p_i}\right]$

26. 0.86,   27. $\frac{2}{3}$ 28. $\frac{1}{2!}-\frac{1}{3!}+\frac{1}{4!}-\frac{1}{5!}\text{.......}\pm\frac{1}{n!}$

Exercise 2(D)

1. $\frac{pp_1}{p_2+p(p_1-p_2)}$ 2. 0.16   3. $\frac{a(c+d)}{a(c+d)+c(a+b)}$ 4. $\frac{5}{8}$ 5. $\frac{1}{3}$

6. $\frac{3}{4}$ 7. $\frac{7}{20}$ 8. $\frac{15}{29}$ 9. $\frac{4p}{3p+1}$ 10. (i) $p_1+p_2+p_3-p_1p_2-p_1p_3-p_2p_3+p_1p_2p_3$

(ii) $p_1p_2+p_1p_3-p_1p_2p_3$ 11.$1-(1-p_1)(1-p_2)......(1-p_n)$ 12. ${}^nc_2\ p^2(1-p)^{n-2}$

(ii) $np^{n-1}(1-p)+p^n$ 13. $\frac{19}{50}$

(iii) $ABC \cup A^cB^cC^c$

12. (i) None of the events happen (ii) Exactly two of the events happen

(iii) At least one of the events A and B happens but C does not happen

13. {c, d, e}, {b, d, e}, {b, c, e}, {b, c, d}, {a, d, e}, {a, c, e}, {a, c, d}, {a, b, e} {a, b, d}, {a, b, c} each with probability 0.1.

15. $\frac{1}{2}$ 16. $\frac{1}{12}$ 17. $\frac{5}{11}$ 18. $\frac{5}{11}$ 19. $\frac{3}{8}$ 20. $\frac{63}{190}$

21. $\frac{7}{12}$ 22. $\frac{2}{9}$ 23. 0.6 24. [0.7, 0.8] 25. $\frac{9}{25}$

26. $\frac{1}{4}$ 28. $\frac{3}{8}$ 29. $\frac{17}{35}$ 30. $\frac{5}{12}$ 31. $\frac{5}{7}$ 32. 0.86

33. $(2p-p^2)$ 34. $3p(1-p), \frac{3}{4}$ 35. 0.8 36. $\frac{a}{a+b}$ 37. $\frac{11}{25}$

38. $\frac{3b+1}{3(a+b+1)}$ 39. $p-2p^2+p^3$

40. $(1-p_1)(1-p_2).....(1-p_n)\sum_{i=1}^{n}\frac{p_i}{(1-p_i)}$ 41. $\frac{5}{9}$ 42. 0.96

43. $np(2+p-np)/2$ 44. $\frac{28}{31}$ 45. $\frac{12}{17}$ 46. $\frac{28}{69}$

P :	$\frac{3}{66}$	$\frac{20}{66}$	$\frac{30}{66}$	$\frac{12}{66}$	$\frac{1}{66}$

3.

X :	3	4	5
P :	0.1	0.3	0.6

4. $(e-1)^{-1}$, $(e+1)/2e$

5.

Z :	$\frac{+4}{-4}$	$\frac{+3}{-3}$	$\frac{+2}{-2}$	$\frac{+1}{-1}$	0
P :	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$

6. (i) 12 (ii) 0.0272, 0.6875 7. (i) $\frac{1}{2}$ (ii) $\frac{1}{2}$ 8. $1\frac{7}{11}$ (ii) $5\frac{3}{11}$

9. $\frac{8}{9}$ 1 0. Rs 5.67p 11. (i) 3.50 (ii) 24.60 12. Rs 2 13. $3(n+1)/2$

14. (i)

Y :	1	2	3	4
P :	.20	.30	.09	.41

(ii) 0.56 (iii) 0.16 15. (B) 16. Rs 7.41p

17. (i) $\frac{(n+1)^2\,(2n+1)}{6}$ (ii) $\frac{(n+1)\,(2n+7)}{6}$

18. (i) $(n+1)$ (ii) $\frac{(n+1)\,(3n+2)}{12}$ (iii) $\frac{(n+1)\,(n-2)}{6}$

19. (i) $\sum a_i\ E(X_i)$ (ii) $\sum a_i^2\ Var(X_i) + \sum_{i\neq}\sum_{j} a_i a_j Cov(X_i X_j)$

20. (i) $m\overline{X}$ (ii) $m(N-m)\ s^2/(N-1)$

3), (2, 4), (2, 5),(3, 4), (3, 5), (4, 5)} each with prob. 0.1

11. {H1, H3, H5, T2, T4, T6} each with prob. $\frac{1}{6}$ 12. $\frac{7}{25}$ 13. $1-\frac{{}^{9}C_4}{{}^{11}C_4}$

15. $\frac{6}{11}$ 16. $\frac{1}{52}$ 17. 0.94 18. [0.75, 1], [0.37, 0.62] 19. (i) T

(ii) F (iii) F (iv) T 21. (i) $p_1p_2(1-p_3)(1-p_4)+p_1p_3(1-p_2)(1-p_4)+p_1p_4(1-p_2)(1-p_3)+p_2p_3(1-p_1)(1-p_4)+p_2p_4(1-p_1)(1-p_3)+p_3p_4(1-p_1)(1-p_2)$

(ii) $p_1p_2p_3(1-p_4)+p_1p_2p_4(1-p_3)+p_1p_3p_4(1-p_2)+p_2p_3p_4(1-p_1)+p_1p_2p_3p_4$

22. $1-a^{n(n+1)/2}$ 23. 10/19 24. 3/7

25. (i)

Z :	3	4	5	6	7	8	9
P :	0.1	0.1	0.2	0.2	0.2	0.1	0.1

(ii)

Z :	1	2	3	4
P :	0.4	0.3	0.2	0.1

26.

Z :	3	4	5	6	7	8	9	11	12	13	14
P :	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

27. 2, 3/2 28. $A^2V(x)+B^2V(y)+2AB\text{Cov}(x, y)$, Cov (x, y) = 0

29. $\frac{1-p}{p}$ 30. 1, 1 31. $acV(X)+bdV(Y)+(ad+bc)\text{ Cov }(X, Y)$

(i) Elementary event (ii) Exclusive event

(iii)Exhaustive event (iv) None of these

(c) Pick up the incorrect statement from below :-

(i) $(A \cap B)^c = A^c \cup B^c$ (ii) $A - B = A \cap B^c$

(iii) $A \cup B = A \cup (A^c B)$ (iv) $A \cap B = (A \cup B)^c$

(d) In three successive tosses of a coin the number of elementary events is :-

(i) 3 (ii) 6 (iii) 8 (iv) 9

(e) The number of ways of arranging the letters of the word MODE so that the vowels occupy the even places is -

(i) 6 (ii) 4 (iii) 8 (iv) None of These

(f) Pick up the incorrect statement from below -

(i) The empty set is a subject of every set.

(ii) Every set is a subject of the universal set

(iii) The universal set and the empty set constitute an algebra

(iv) If P (A) = 0 then A is the empty set.

(g) The number of arrangements of the letters of the word 'NOTION' is -

(i) 720 (ii) 180 (iii) 120 (iv) 48

(h) Pick up the incorrect statement from below -

If A and B are independent events then-

(i) B and A are independent

(ii) B^c and A are independent

(iii) $P(A \mid B) = P(B \mid A)$

(iv) $P(A^c \mid B) = P(A^c)$

(iii) If $P(ABC) = P(A)P(B)P(C)$, then A, B, C are mutually independent.

(iv) None of the above statements is true.

(j) The statements below relate to Bayes theorem :

(i) Bayes theorem is a formula which computes a conditional probability.

(ii) Bayes theorem is used to compute the probability of an earlier event on the basis of a subsequent observation.

(iii) The posterior probability computed by Bayes theorem supersedes prior probability.

Mark your answer with code

(i) If only I & II are true

(ii) If only I & III are true

(iii) If only II & III are true

(iv) If all the statements are true

2. **Fill in the blanks :**

(a) Classical definition of probability is applicable only if the elementary events are ________.

(b) The symbolic representation of the occurrence of only one of two events A and B is ________.

(c) The symbolic representation of the occurrence of at least two of 3 events A, B, C is ________.

(d) The correct inequality between P(A), P(B) and $P(A \cap B)$ is ________.

(e) In simultaneous toss of two symmetric dice, the probability of getting a total score exceeding 10 is ________.

(f) The number of conditions required for the independence of three events A, B, C is ________.

(i) If (x,y) represent a pair of numbers selected from the set {1, 2, 3, 4}, then E(XY) is equal to ________.

3. **Very short answer type :**

(Answer the following questions in one or two words or in one sentence):-

(a) Give an example of three mutually exclusive and exhaustive sets from the universal set {1, 2, 3,.......,8, 9, 10}.

(b) Write $A \cup B \cup C$ as the union of three disjoint sets .

(c) Which subset is a subset of every subset ?

(d) In how many ways can the letters of the word 'TWELVE' be permuted so that the letters T and L do not occur together ?

(e) From a group of 5 boys and 4 girls; in how many ways can a group of one boy and one girl be formed ?

(f) Write down an inequality between P(A), P(B) and $P(A \cup B)$.

(g) Define independence of 3 events A, B, C.

(h) 3 numbers are randomly selected from the set {1, 2, 3, 4, 5, 6, 7}. What is the probability that all of them are ≤ 5.

(i) If X is the score in random toss of a symmetric die, find $E(X^2)$.

(j) Under what conditions, var(X+Y) = var (X) + var (Y) ?

(k) (x,y) represent the numbers obtained by selecting X and Y from the set {1, 2} with replacement. Write down the bivariate distribution of (X, Y).

(l) Under what conditions, var (X+Y) = var (X−Y) ?

(g) $1-(1-P)^3$ (h) $0.5 \le P(A \cup B) \le 0.9, 0 \le P(AB) \le 0.4$ (i) $5\frac{5}{6}$

★★★

3A.1 HISTORY, MEANING AND SCOPE OF STATISTICS

History reveals that Statistics as a subject has originated since ancient times in the form of "Science of State Craft". It was a by-product of the administrative activities of the state. The word 'Statistics' seems to have its origin in the Latin word 'Status' or the German word 'Statistik' or the French word 'Statistique', each meaning a political state. In ancient times it was the function of the government to collect data regarding the 'population' and 'property and wealth' of the state for framing military and fiscal policies. The efforts of the state for such collection of data was aimed at assessing the manpower of the country with a view to safeguarding it against external aggression and providing a basis for introducing new taxes and other levies.

In India, an efficient system of collection of official and administrative statistics existed even 2000 years ago. Historical evidences about the prevalence of a very good system of collecting vital statistics and registration of births and deaths before 300 B.C. are available in 'Artha shastra' written by Kautilya. The records of land, agriculture and wealth statistics maintained during the reign of emperor Akbar (1556 – 1605 AD) are evident from the book 'Ain-e-Akbari' written by Abdul Fazl (1596-97 AD), one of the nine gems of Akbar.

With the passage of time such information was considered inadequate. Besides births and deaths, other details like marriages, divorces, emigrations, immigrations etc. were required for the smooth administration of the state. The importance of such vital statistics was first conceived by John Graunt of London and was later developed by Casper Newman, Sir William Petty, Dr. Price etc. Subsequently these information, called data, became the basis of the study of various branches of science like Agriculture, Biology, Medicine, Economics, Political Science etc.

In the mid-seventeenth century, the concept of statistics was re-oriented with the introduction of "Theory of Probability" which linked up the theory of Statistics with in the field of modern Statistics are Karl Pearson (1857-1936) W.S. Gosset etc. Some of the Indian statisticians who have significant contributions to the field of Statistics are P. C. Mahalanobis, C.R. Rao and others.

3A.2 DEFINITION OF STATISTICS

The word statistics is being used to convey two meanings. Statistics in plural form refer to statistical data and in singular form refers to principles and techniques used in the collection, analysis and interpretation of statistical data. It is not possible to include all the definitions of statistics given by different persons, considered both as singular and as plural, within the scope of this book. However, we include only one definition of each form with our comments.

3A2.1 Statistics as Statistical Data

Prof. Horace Secrist's definition of Statistics is considered the most exhaustive one as it clearly points out all the aspects of the subject. In the words of Secrist, Statistics may be defined as "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".

Comments

i. **Statistics are aggregate of Facts :** Single or isolated figures cannot be called Statistics unless they are aggregate of facts or a part of aggregate of facts relating to any particular sphere of enquiry. For example, the marks of a student in an examination or the income of a particular person or the price of a pen would not constitute statistics as such figures are not related and cannot be compared. On the other hand, the marks of the students of a class in an examination, the incomes of all the employees of an organization or the price of a product over different time periods constitute Statistics.

However, in the presence of the joint effect of a number of forces acting on a single item, the effects of some groups of factors can be studied through statistical techniques.

iii. **Statistics are Numerically Expressed :** Only numerical statements of facts constitute statistics. Qualitative statements like 'the production of paddy crops in Odisha is decreasing' or 'the production of sugar is not sufficient' or 'the standard of living of slum dwellers is increasing' do not constitute statistics. Such statements are vague and imprecise. On the other hand, statements like 'the estimated production of steel by SAIL (Steel Authority of India Limited) in the next five years would be 520 million tons' is a statistical statement.

iv. **Statistics are Enumerated Expressed or Estimated According to Reasonable Standards of Accuracy :** Numerical data pertaining to any sphere of enquiry can be obtained in two ways viz. (a) by cent percent counting (or measurements) called complete enumeration (census survey) or (b) by estimates through partial counting (or measurement) called sampling. In complete enumerations, data are expected to be exact and accurate while estimates obtained through samples are not expected to be as precise and accurate. On many occasions, complete enumeration is either inconvenient or impossible. In such situations one has to be contented with estimates which are obtained through samples. As the estimated values are usually not precise and accurate, while collecting data, reasonable standards of accuracy must be maintained. The degree of accuracy depends largely on the nature and purpose of the enquiry. For example, while weighing vegetables, weights of some grams may be ignored but while weighing precious metals like gold, even fractions of milligrams cannot be ignored. Similarly, while measuring the

v. **Statistics are collected in a systematic manner :** Before collection of statistics, a suitable planning of data collection should be made. Then data should be collected in a systematic manner. Data collected unsystematically in a haphazard manner may lead to fallacious and absurd conclusions.

vi. **Statistics are collected for a predetermined purpose :** The purpose of the enquiry must be defined in clear and concrete terms in advance and data should be collected keeping in view the purpose of the enquiry. Too many data, some of which will never be examined and analysed, should not be collected. Only data consistent with the purpose of enquiry need be collected. Care must be taken to ensure that no essential data are omitted. For example, while studying for the cost of living of the lower income group people of Mumbai, one should select only those commodities or items which are actually consumed or used by the lower income group people of Mumbai.

vii. **Statistics should be placed in relation each other :** In order that numerical facts (data) be called statistics, they should be comparable. Statistical data are usually compared with respect to time (period) or regions (place). For example, the population of India at 2001 may be compared with those of earlier years or with the population of other countries viz, China, USA etc. Valid comparison is possible only for homogeneous data i.e. data relating to the same phenomena or subjects. But data relating to the height of an individual and his income do not constitute statistics and cannot be compared.

In the absence of the above characteristics, numerical facts cannot be called statistics. Hence, all statistics are numerical facts but all numerical facts or statements are not statistics.

and interpretation of numerical data."

Comments

This definition of statistics is very simple, concise and most exhaustive. It includes all the four stages in statistical investigations viz. collection of data, presentation of data, analysis and interpretation of data.

It may therefore be concluded that the science of statistics is a study of methods applied in collecting, analysing and interpreting quantitative data affected by multiple causation; in any field of enquiry.

3A.3 SCOPE AND LIMITATIONS OF STATISTICS

In the ancient times when Statistics was considered as the science of 'State-craft', it was used by governments to collect information relating to manpower, property and wealth for devising military and fiscal policies. But with the passage of time the concept of 'State Welfare' was considered almost all over the world. As a result, the scope of statistics widened to social and economic phenomena. Further, with the development of statistical techniques, now a days Statistics is viewed not only as a mere device of collection of numerical data but also as a sound technique to collect, process, analyse and draw valid conclusion from them. As a result, its application extends to all branches of science – social, physical and natural. It is finding application in diversified fields like economics, business, industry, management, agriculture, commerce, education, psychology, sociology, biological sciences, medicine etc. It is rather impossible to think of any sphere of human activity where statistics is not used. It will not be out of the way to state that the modern culture has become a statistical culture.

We discuss, in brief, a few fields in which statistics is applied.

of immense help in promoting human welfare. All ministries and departments of the government depend heavily on factual data for their efficient functioning. For example, in the Planning Department, Statistical data and Statistical methods are indispensable to the Government for planning of future economic programme. The study of population movement i.e. population estimates, population projection, sex wise birth and death statistics, age and sex distribution, provide fundamental tools for overall planning and evaluation of economic and social development programmes.

The use of statistical data and statistical methods are so wide in government functioning that almost all ministries and departments of the government have separate statistical units. In most countries, the state is the single unit which is the biggest collector and user of statistical data. The main statistical agencies in India are Central Statistical Organisation (CSO), National Sample Survey Organisation (NSSO) and the Registrar General of India (RGI). They collect data at a national level periodically.

3A.3.2 Statistics and Economics

Sir William Petty in his book named 'Political Arithmetic' published in 1690 has mentioned the importance of Statistics in Economics. But, in fact, Statistics and Economics became closely associated only in the recent past. In the earlier stages Economic Theories were based on deductive logic only. Economists never used quantitative data for development of economic doctrines. Besides, in those days figures were considered as life less, rude and coarse and hence avoided. Gradually, Economists realized that theory must be based on the reality of life and facts. So, the use of statistics gained popularity among Economists to derive economic theories through inductive process. In the present scenario, Statistics and Economics have become so intimate that in 1890, Prof. Alferd Marshall observed and commented 'Statistics are the straw out of which I, like other economists, have to make bricks.'

demand, statistics of consumption helps in finding the way in which people of different income groups spend their income. Statistics plays an important role in the study of prices, exchanges, computation of National Income, distribution of income and wealth, growth of population, unemployment, poverty etc. Econometrics, now used in economic research, comprises of the application of statistical methods to the theoretical economic methods.

3A.3.3 Statistics and Business Management

Now a days statistics is widely used in business concerns. The business men and the business organisations depend heavily on Statistics and Statistical methods at every stage of their activities. Management of business has become smooth and easy by the application of statistics. Before the Industrial Revolution, business was in its early stage. Business people were taking decisions on the basis of old methods like hit or miss or leaving the future to chance. But after the industrial revolution, there has been rapid development of trade, commerce and industry. The business managements try to expand their business while the future is uncertain. They find themselves helpless on the faces of uncertainty. They have realised that success of a particular business depends on accuracy, propriety and precision of statistical data analysis. Business executives feel the importance of statistics while going to promote a new product or looking for expansion of the existing ones. For making a right decision they depend on statistics by considering data relating to price of raw materials, price and demand of similar available products in the market, problem of manpower etc.

Statistical data are used by business enterprises for business forecasting and quality protection of the products. It is also used in the sector of life insurance. In short, it can be said that statistics plays a very important and essential role in all fields of business activity.

statistics to physical sciences came late. Now a days, statistics is used in Astronomy, Chemistry, Engineering, Geology and in certain branches of Physics.

3A.3.5 Statistics and other uses

Statistics and Mathematics are intimately related. In a sense, it can be said that the modern Statistics is an off-shoot of Mathematics. Mathematical Statistics is the outcome of increasing role of mathematics in to statistics. Mathematics has also freely used methods of Statistics. When analytical methods prove inadequate, mathematicians use numerical methods from the purview of Statistics. Statistics is frequently used in Computer Science, social sciences like Education, Psychology, Sociology etc. for experiments and drawing valid conclusions by considering and analysing data obtained through experiments.

3A.4 LIMITATIONS OF STATISTICS

We have seen how statistics is indispensable in the study of all spheres of human activities. But still, like all branches of study, it has its own limitations. We enumerate below some of these limitations.

1. Statistics does not suit to the study of qualitative phenomena

As the definition suggests, statistics are numerical facts. Hence it is useful in the study of those objects of enquiry which admit quantitative measurements. Accordingly, objects signifying qualities like honesty, intelligence, poverty, justice, culture, nationality, beauty etc. do not come under the purview of statistics. But statistical techniques may be applied indirectly to those objects by first reducing the qualitative expressions to precise quantitative measurements. For example, the intelligence of a group of students can be expressed in terms of Test Scores and then analysis can be carried out by statistical methods.

statistical study. On the other hand, if we consider the marks of all the students or a section of students of a class in a particular subject in an examination or the profits of all the business organisations of a city or the production of a group of business houses etc; these will fall within the purview of a statistical study.

3. Statistical laws are true only on an average

Statistical laws are not as accurate as laws of natural or physical sciences. They are not precise or correct universally. Statistical laws hold good only on an average. Statistics deals with such phenomena which are affected to a marked extent by multiplicity of causes and it is not possible to study the effects of each of those causes separately like experimental methods. Because of this limitation in statistical methods the conclusion drawn for one group may not agree with that of another similar group.

4. Statistics is liable to be misused

The use of statistical methods by inexperienced and untrained persons might lead to fallacious conclusions. As such, statistics should be used by experts in the field. Statistical methods are dangerous tools in the hands of inexperts. Since statistics are numerical figures, they can be manipulated by dishonest and unskilled persons. Rightly, W.I King says "Statistics are like clay of which you can make a God or a Devil as you please." So, we can conclude that Statistics is of immense help and is of great value to those who understand its proper use.

3A.5 DISTRUST OF STATISTICS

Distrust means lack of confidence or disbelief. By distrust of statistics one carries the impression of lack of confidence in Statistics and Statistical methods. To the misfortune of Statistics and Statistical methods, their popularity has been adversely affected by people who have described statistics with various comments, some of which are stated below.

"An ounce of truth will produce tons of statistics."

(i) Since statistics deals with figures, they are believed easily. The facts based on figures are convincing. But, unfortunately, they do not always bear on their face the hall mark of their quality. So, reliable and unreliable data look alike.

(ii) To establish certain results which are not true, inaccurate figures or incomplete data can be used.

(iii) Truth can be concealed by manipulation.

We state below some examples to illustrate how fallacious conclusions can be drawn from statements.

(a) In the financial year 2006-07 the profits of firm A is Rs. 2 lacs while that of firm B is Rs.3 lacs. So firm B is decidedly better than firm A.

(b) 80 percent of the people who drink alcohol die before attaining the age 90. So, drinking alcohol is harmful for longevity.

(c) The average scores of two students, say, Priya and Rupa, are equal. So Priya and Rupa have the same standard.

(d) The number of road accidents by women drivers is less than those of the men drivers. So women drivers are more competent than their male counterparts.

Although the above statements appear to be true at the out set, on a close observation one will discover that these can be false.

Exercises – 3A.1

1. Define Statistics as numerical data and discuss its various features.
2. Define Statistics as Statistical methods and give your comments.
3. Discuss the importance and scope of Statistics.

may be concluded that smoking causes cancer.

II. The increase in the price of a commodity was 10% in 2000. Then the price decreased by 15% in 2001 and again increased by 5% in 2002. So the net increase in the price is 10 – 15 + 5= 0.

III. The average output in a factory was 2500 units in January 2005 and 2400 units in February 2005. So the workers were more efficient in January than in February 2005.

IV. The percentages of success of girls was 80% while those of boys was 70%. So girl students are more intelligent than boys.

V. The average monthly income of two families are Rs.5000/- each. So the standard of livings of both the families are the same.

3A.6 COLLECTION OF DATA

From earlier section we know that the plural concept of statistics refers to data. Data are numerical facts obtained either by counting or by actual measurements of units or objects or observations under study. The person who actually counts or measures is called 'investigator' and the persons from whom information is collected (or numerical measurements are taken) are called 'respondents'. Collection of data is the first step in any statistical investigation (or enquiry). As such, utmost care should be taken for collection of data. Statistical data are classified in to two categories viz. Primary data and Secondary data. Primary data are first hand information i.e. those which are collected for a specific purpose directly from the field of enquiry and thus are original in nature while Secondary data are those which have been collected by some other person or agency or organisation and have already undergone statistical processing at least once. Primary data may be considered as raw materials to which statistical methods are applied to give a final shape.

by the Revenue Department of the Government for assessment of rent and cess are primary data for the Revenue department but if these data or a part of these data are used for calculation of National Income, those become secondary data. Hence, on a closer observation, it will be obvious that the difference between primary data and secondary data is one of degree only.

3A.6.1 Collection of Primary Data

Primary data are collected either by cent percent counting (or measurement) called 'complete enumeration or census survey' or by counting a representative part, called 'sampling or sample survey'. Collection of primary data is a labourious, time consuming and expensive process and so should be taken up by well organised bodies. Before collection of primary data, the following preliminaries must be examined carefully.

1. The purpose of data collection.
2. The target population i.e. the population where from the data has to be collected.
3. The units or objects on which measurements are to be taken.
4. The body or the organisation or the person who would collect data.
5. The type of survey for collection of data.
6. The degree of accuracy desired.

We discuss, in brief, the above preliminary steps as follows :

1. The purpose of the enquiry should be specified in clear and concrete terms. This will enable the investigator to decide about the nature and scope of data to be collected and the statistical methods to be used for analysis.
2. Decision about the population (the target population) where from data would be collected has to be made. The target population is one that gives the desired information about the enquiry.

(investigators) should be recruited on the basis of knowledge, work experience, interest for the job etc. The investigators selected should be honest and efficient. Sometimes, instead of directly appointing the investigators, the job of collection of data may be entrusted with dependable agencies or organisations having interest in the job.

5. The type of survey for collection of data is to be decided i.e. decision is to be taken as to whether complete enumeration method or sampling method would be appropriate. The choice is to be made keeping in view the availability of time, resources (cost and manpower) and the degree of accuracy. In case, information regarding each unit of the population is desired, complete enumeration is the only choice.

6. Degree of accuracy desired must be specified so that the investigators would keep themselves alert while making actual measurements or counting.

3A.6.2 Methods of Collection of Primary Data

Following methods are commonly used in the collection of primary data.

i) Direct personal observation.

ii) Indirect oral investigation

iii) Schedules sent through investigators.

iv) Questionnaires sent by mail

v) Entrusting the data collection to a Local Agency or Correspondents.

3A.6.2(a) Direct personal observation

In this method the investigator moves directly to the spot of enquiry and collects the required information from the informants or respondents by meeting the people and conducting the enquiry and observing the facts personally. In such enquiry the investigator enquiry is narrow. But for extensive study, this method should not be used as it will involve more time, cost and manpower. The data obtained by this method are generally more dependable and accurate provided the investigators collect the data honestly and seriously with care and accuracy and do not fabricate data without going to the spot. Correct information can be extracted by the investigator by cross questioning the respondent when the investigator doubts the integrity of the respondent or he feels that the respondent is exaggerating or hiding the truth. This method works well when the investigator is skilful and capable of handling sensitive questions. As the method is subjective in nature, the success depends purely on the wit, skill, tactics, insight, diplomacy and courage of the investigator. The personal bias, prejudices and whims of the investigator may influence the data in some cases which may adversely affect the data.

3A.6.2.(b) Indirect Oral Investigation :

On many occasions, the persons who are to furnish the required information are either unwilling to take part or reluctant to co-operate or the sphere of enquiry is very wide or the source of information does not exist - (i.e. has been destroyed). In such cases, the required information can be collected indirectly by a method called Indirect Oral Investigaton (or witness method). In this method, factual data relating to the enquiry are collected by interviewing persons who are directly or indirectly concerned with the subject matter and possess the required information. Such persons are called 'witnesses or informers'. For example, to collect information on some social evil like addiction to gambling, drinking alcohol etc. the persons directly involved would not give the correct information. In such cases, friends, relatives, neighbours or other family members may be contacted. The investigator prepares a small list of questions relating to the subject matter of investigation and asks those questions to the witnesses and records their replies. This method of collection of factual data are usually adopted by most of the commissions and committees

(i) should know full facts about the problem under study.

(ii) should not be prejudiced and biased.

(iii) should not give colour to facts

(iv) should not lie or misguide the investigator.

(v) should be capable of expressing themselves correctly and

(vi) should not be inherently an optimist or pessimist.

Merits and Demerits

In this method, the investigator can exercise his skill, intelligence, tact etc. in extracting the correct information by cross examination. This method is less expensive and less time consuming in comparison with the method of direct personal observation. Investigators can be selected by the suggestion and views of experts. But the success of the method depends on the integrity, insight, skill, intelligence and efficiency of the investigator. The investigators are to be properly trained to collect factual information. A wrong and improper choice of the witness might affect adversely and spoil the entire purpose of the study.

3A.6.2.(c) Schedules sent through investigators

A schedule is a sheet of paper containing a list of questions in the shape of a proforma which is filled in by the investigator or the enumerator (field agent) in a face to face interview with the respondent. The investigator visits the respondent personally and records his replies to various questions listed earlier on a sheet of paper or a proforma, either printed or cyclostyled. This method is usually adopted by big business concerns, public enterprises, research institutions or Government. The population census in India, held in an interval of every ten years, is conducted by adopting this method for collection of data.

extensive study. The method can be effectively carried out even if the respondent is illiterate.

It is a fairly expensive and time consuming process and involves more man power. The success of this method depends on the skill, integrity, efficiency of the investigator and the nature of the questions. Proper training must be given to the investigators to acquaint themselves with the purpose and scope of the study and to explain them clearly to understand the implications of various terms used so that they can clarify the doubts of the respondents, if necessary. The investigators should be unbiased and should not twist the answers and should not impose their opinions on the respondents.

3A.6.2(d) Mailed Questionnaire method

A questionnaire is a printed or a cyclostyled form containing the list of questions relating to a particular field of enquiry and sent to the respondent by mail, answers of which are to be recorded by the respondents and sent back by mail. The difference between a schedule and a questionnaire is that, a schedule is filled in by the enumerator while a questionnaire is filled in by the respondent. When the answers to the list of questions printed on the form or on several forms after being recorded by the respondents themselves are collected by the investigator, it is called collection of data by mailed questionnaire method.

In this method, a covering letter written in polite language is sent to the respondent along with the printed list of questions. The covering letter should explain the aim and objective of the study and the importance of the respondent's participation and co-operation in the investigation. The respondent should be promised that his answers would not be divulged in case he wanted so. A prepaid addressed envelope should also be sent along with the questionnaire so that the respondent can send back the filled in set of questions at the investigator's expenses. The respondent may also be promised to be provided with a gist of the final report of the study if he desires so. If possible, some gifts in the form of coupons may be sent to the respondent so that he takes interest to fill in and return.

This method is effective only for a literate population and can't be used for an illiterate population. The success of this method depends on the co-operation of the respondents and timely reply. In this method, the response is usually low i.e. sometimes the respondents may not respond at all or respond haphazardly and incompletely or provide wrong and motivated answers deliberately. They may suppress some facts and exaggerate some other facts. They may be unwilling to provide answers in writing. In such cases the purpose of the study is lost. Since the only correspondence between the respondent and the investigator is the questionnaire, there is no scope to clarify the respondent's doubts, if any, or explaining the meaning of different expressions used in the questionnaire. This method is thus less reliable in comparison with the method of schedules sent through investigators.

3A.6.2(e) Local Agency or correspondent

This method consists of collecting primary data through agents or by local correspondents who collect data in their own fashion and to their own likings. They submit their reports periodically to the central office for processing and analysis. This method of data collection is usually adopted by news paper and periodical agencies who require information relating to different fields like economic trend, stock market, accidents, sports, riots, strikes etc. This method is adopted by the Government for obtaining estimates of agricultural production.

Merits and Demerits

It has the advantage of being least expensive. It is applicable for extensive investigation over a variety of activities.

The data collected through agencies or correspondents can not be very reliable. So this method is used in those cases where the purpose of the investigation is only to get a rough estimate. This method should not be used where a high degree of precision is desired.

These are :

1. Number of questions should not be large. Only questions relating to the essential points of the enquiry should be included. A reasonable number of questions, which a good questionnaire contains, should be 15 to 25.
2. The questions should be simple, short, clear, unambiguous, non-offending, courteous in tone (words) and just to the point.
3. As far as possible, the questions should be capable of objective answers i.e. such which can be answered by 'Yes' or 'No' or by ticking the boxes or by putting cross marks or indicating one of the several alternative answers. If possible, lengthy questions should be split up to facilitate answering.
4. Questions of personal nature (such as income, extent of property, offences committed etc) or those which hurt the feelings and sentiments of the respondents should not be asked.
5. Questions whose answers need calculations should be avoided.
6. Vague questions and words having multiple meaning should not be used.
7. Questions should be arranged in a natural and logical sequence i.e. questions of the same category should be put together.
8. There should be some questions which are corroboratory in nature. i.e. some questions should be there to cross check the answers of the respondents.
9. Questions should be pre-tested i.e. should be tried on a small group of respondents of the population to sort out the draw backs of the questionnaire. This would help the investigator in finding out whether there are duplication of the questions or some necessary questions have been omitted or vague and in appropriate questions have been included so that necessary steps can be taken to rectify the omissions and commissions. From pretest, the investigator

DEPARTMENT OF STATISTICS

.... UNIVERSITY, ODISHA

To

Mrs/Mr/Ms. ....

....

....

Dear Madam/Sir,

A survey is being taken up by this Department to study the effect of advertisements of textile goods on consumers. You must be aware of the fact that, now a days, the textile companies are spending a lot of money on advertisements which is finally borne by the consumers. Through the present survey we attempt to study whether it is worth spending so much on advertisements. The information provided by you will be used for the purpose of this survey only and will be treated confidential. If you so like, the gist of the final findings of this survey will be supplied to you free of cost.

It would be a great favour if you kindly fill in the enclosed questionnaire and return the same by post at the earliest. An addressed and stamped envelope is enclosed herewith for your convenience.

Thanking you for your co-operation.

Yours sincerely

....

Encl : 1. The questionnaire

2. A stamped envelope.

A. Respondent's profile :

1. Name (in capital letters) ....

2. Address ....

....

3. Age in complete years ....

4. Sex (put ✓ mark in the appropriate box) Male □ Female □

5. Educational Qualification

(Put ✓ mark in the appropriate box)

Non – matric □

Matric □

Graduate □

Technically Qualified □

6. Occupation

(Put✓ mark in the appropriate box)

Agriculture □

Service □

Business □

7. Annual Income

(Put✓ mark in the appropriate box)

Below Rs.50,000 □

Rs.50,000 – Rs.1,00,000 □

Rs.1,00,000 – Rs.5,00,000 □

More than Rs.5,00,000 □

8. No. of members in the family Male □ Female □

(Fill the boxes using numbers)

2. Whether you have some prior knowledge about the

(strike out the one not applicable)

(i) Quality	Yes/No
(ii) Price	Yes/No
(iii) Brand	Yes/No

3. If your answer to any of the above in (2) is Yes, then how did you know about this?

(strike out the one not applicable)

(i) Through advertisements

(ii) Seeing some one else

(iii) Through friends

(iv) Door canvassing

4. After entering into a textile shop you start enquiring about

(Put ✓ mark in the appropriate box)

	Mostly	Sometime	Never
(i) Textile of a particular brand	□	□	□
(ii) Textile in a certain price range	□	□	□
(iii) Textile based on quality	□	□	□
(iv) Textile seen in the advertisement.	□	□	□

5. The factors influencing your purchasing decision are

(Rank in order of preference by writing 1,2,3 etc.)

(i) Price	□
(ii) Quality	□
(iii) Colour and design	□
(iv) Impact of company advertisement	□
(v) Advice of the seller	□
(vi) Friend's or relative's advice	□
(vii) Display in the shop bearing company's name	□

(iii) Children's wears □

(iv) For self □

(v) For spouse □

(vi) For gift □

2. Do you think that textile advertisements seen by you are helpful in your decision making ?

(Strikeout the one not applicable) Yes/No

3. Do you think that textile advertisements do not help you in your decision making?

(Strikeout the one not applicable) Yes/No

4. If the answer to No.3 above is Yes, then the reason you feel could be _

(Strike out the one not applicable)

Advertisements :

(i) Highlight the company's name Yes/No

(ii) Do not specify the quality Yes/No

(iii) Do not indicate the price range Yes/No

(iv) Do not mention the places where the advertised cloth is available Yes/No

(v) All companies do not advertise Yes/No

D. Source of information about the company's products

(Put ✓ mark in the appropriate boxes)

1. How did you know about the company's name?

(i) Visiting shops □

(ii) Through associates □

(iii) Through advertisements □

E. Your suggestions : ....

1. What factors you consider helpful to be incorporated in the advertisements?

   (Rank your preferences in order by writing 1, 2, 3, etc)

   (i) The quality range □

   (ii) The price range □

   (iii) The colours and designs □

2. How often the advertisements should appear? Regularly □

   (Put ✓ mark in the appropriate box) Occasionally □

3. Where should it be advertised?

   (Put ✓ mark in the appropriate box / boxes)

   (i) News papers □

   (ii) Pamplets □

   (iii) Media □

   (iv) Retail shops □

4. Any other suggestion (limit to 25 words)

....

....

Date : ....   Signature of the respondent

3A.8.1 Published Sources

There are a number of Government, semi-government, autonomous and private bodies which collect data relating to prices, production, income, business, industry and socio-economic phenomena regularly or occasionally. These collected data are processed and published monthly, quarterly or annually. Some of the Central Government publications are :

(i) Monthly Abstract of Statistics published on a monthly basis containing data of production of selected Industries in India.

(ii) Sample Surveys of Current Interest in India published annually.

(iii) Census data and census reports published decennially.

Some publications of Semi-Government Organisations are :

(i) Data relating to Annual Report of Banks published annually by Reserve Bank of India (RBI)

(ii) Data relating to Currency and Finance published by RBI on a monthly basis.

Besides, various research institutes, individual researchers, different departments of various universities and statistical organisations collect and publish data.

Reports of various committees and commissions appointed by the Government provide published data. For example, Kothari Commission Report on Educational Reforms, Lindo Commission Report on conduct of Elections in Educational Institutions or Pay Commission Reports etc.

News papers, International Publications by bodies like IMO, UNO, WHO etc. are some of the sources of secondary data.

Some private sectors like Chambers of Commerce also publish data on a regular basis.

3A.9 PRECAUTIONS IN THE USE OF SECONDARY DATA

Published statistics should not be taken for granted at their face value. Scrutiny of secondary data is necessary for examining its accuracy, suitability and adequacy. So, before using any published or unpublished source of secondary data, it is necessary to scrutinise those to ascertain whether they possess the following attributes.

1. **They should be reliable.** Reliability of data can be tested by finding out whether the data

   (a) have been collected by dependable persons and from dependable source

   (b) refer to a normal period of time, comparable to the period of study.

   (c) have been collected by using proper statistical methods

   (d) do not contain deliberate and unconscious bias by the compiler, and

   (e) have the desired degree of accuracy.

2. **They should be suitable for investigation and comparison.** Even if the data are reliable, those should not be used unless considered suitable for the purpose of enquiry. For example, if for the purpose of original investigation, data have been collected by considering one house hold as one unit and for the present purpose, one unit refers to one individual, such data are not suitable for the present purpose. Again, the original source might have collected data which are wholesale prices while for the present purpose retail prices might be required. Further, the purpose, nature and scope of enquiry might be different. In such cases, the secondary data would be unsuitable for the investigation under study.

3. **They should be adequate.** Secondary data although found reliable and suitable may not be adequate for the purpose of investigation. The original source referred

Hence it is very risky to use secondary data collected by other persons unless they have been thoroughly scrutinized and edited.

Exercises – 3A.2

1. Distinguish between (a) primary data and secondary data (b) census survey and sampling methods.
2. Discuss the various methods of collecting primary data.
3. Describe briefly the questionnaire method of collecting primary data and write its advantages and disadvantages.
4. Discuss the essentials of a good questionnaire. What is the purpose of 'covering letter' which is sent along with the questionnaire?
5. Distinguish between schedules and questionnaire methods of conducting survey and write their advantages and disadvantages.
6. What are the different methods of collecting primary data? Why are 'Direct Personal Observations' usually preferred to "Questionnaire" method ? Under what conditions may a Questionnaire method prove as satisfactory as the method of Direct Personal Observations?
7. Explain the method of Indirect Oral Investigations of collecting primary data and indicate its applications.
8. What do you mean by 'Witness method' of collecting primary data? Write its advantages.
9. Explain, with suitable examples, what you mean by secondary data.
10. Describe the various sources of secondary data.
11. "It is never safe to take published statistics at their face value without knowing their meaning and limitations." Elucidate this statement.
12. Explain the various points you would consider before using published statistics. Give suitable examples where ever possible.

(ii) The sources of secondary data are .... and ....

(iii) .... data should be used after careful scrutiny.

(iv) Data are classified as .... and ....

(v) Mostly committees and commissions adopt .... method of collecting data.

(vi) Annual Statistics published by Directorate of Bureau of Statistics and Economics, Govt. of Odisha are called .... data.

(vii) In India, the population census is conducted by adopting .... method of collecting data.

(viii) Collection of data by Direct Personal Observation is suitable for .... study.

(ix) Now a days the elections in various educational institutions are held according to .... recommendations.

(x) Data collected by a research scholar are .... for him.

16. Indicate whether the following statements are true (T) or false (F).

(i) Secondary data are used in those cases where the primary data do not provide an adequate basis for analysis (T/F)

(ii) Secondary data do not require much scrutiny and should be accepted at its face value. (T/F)

(iii) A covering letter sent along with the questionnaire explains the purpose of the investigation. (T/F)

(iv) Witness method of collecting primary data is most suitable for research scholars. (T/F)

(v) There is no difference between the methods of 'schedules sent through enumerators' and 'direct personal observation' of collecting primary data. (T/F) purpose of study. Classification may be defined as "the process of arranging data in groups or classes according to resembiances and similarities." Sorting out facts on the basis of a single characteristic like height, weight or marks is called 'simple classification' and on the basis of more than one characteristics like height and weight, marks in two or more subjects etc. is called 'cross classification'. Thus, classification is the first step of analysis and interpretation of statistical data. In some cases, classification may give such a clear picture of the significance of the material that further analysis may not be required. In case, secondary data are used, they need be rearranged in to different groups having common characteristics to suit to the purpose of the study. For example, the results of 500 successful students in an examination may be grouped as students getting first division, second division, third division and compartmental categories. Sorting out the letters in a sub-post office according to common destinations is another example of classification. The product of a textile manufacturer may be classified according to their sizes viz. small, medium, large, extra large etc.

3B.1.1 Objects of classification

The objects of classification may briefly be stated as follows:

(i) It condenses the mass of data.

(ii) It facilitates comparison.

(iii) It pinpoints the most significant features of the data at a glance.

(iv) It gives prominence to important information and eliminates unnecessary details.

(v) It facilitates statistical treatment of the data.

in two groups 'tall' and 'short', it should be clearly specified as to who would be called tall and who short.

(ii) **It should be exhaustive and mutually exclusive.** The classes should be such that they accommodate all the units or observations and no unit is left out. A good classification should be free from a class like residual or other or miscellaneous class. Such classes do not reveal the characteristics of the data completely. The classes should be non-overlapping so that there is no scope of inclusion of one observation in more than one class.

(iii) **It should be stable**. For meaningful comparison of the results, once a particular pattern of classification is followed, the same pattern of classification must be adopted all through out the analysis and also for further investigation on the same subject.

(iv) **It should suit to the purpose of the investigation**. The classes should be formulated keeping in view the objective of the study. For example, for the comparison of marks of students in a class from time to time, the classification should be by marks and not by the ages.

(v) **It should be flexible**. An ideal classification should be flexible in the sense that it should provide scope to accommodate and adjust to new situations and circumstances.

3B.1.3 Types of classification

The data can be broadly classified in to the following four categories.

(i) Geographical i.e. area or region wise viz. cities, districts etc.

geographical or locational differences between the various items in the data like States, Cities, Regions, Zones, Areas etc. For example, state wise agricultural production in India may be presented in the following manner.

STATEWISE ESTIMATES OF AGRICULTURAL PRODUCTION

Name of the State	Total Production ('000 tonnes)
Andra Pradesh	19,584.4
Bihar	11,396.0
Odisha	12,814.8
Uttar Pradesh	31,862.0
West Bengal	21,148.7
Total	1,06,805.9

Geographical classifications are usually listed in alphabetical order for easy reference. This may also be listed by size to emphasize the area. Usually the first approach is followed in the Geographical Classification.

(ii) Chronological classification.

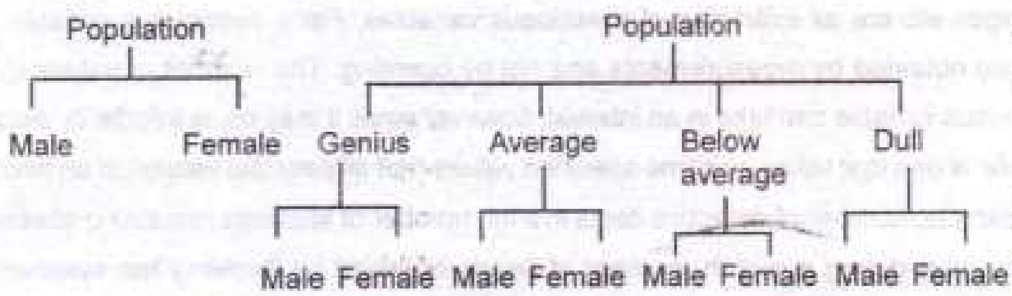
In chronological classification, the data are classified on the basis of difference in time. For example, the population of a country in different years, the production of an industrial concern in different months, the profits of a business organisation in different years etc may be presented as chronological classification. Following example will be instructive.

1951	36.11
1961	43.92
1971	54.82
1981	68.33
1991	84.63
2001	102.86
2011	121.19

(III) Qualitative Classification

When data are classified according to some attributes or characters or quality such as sex, literacy, religion, occupation, beauty etc. the classification is called qualitative or descriptive or according to attributes. Here the attribute under study cannot be measured quantitatively except that one can identify whether it is present or absent in the units of the population under study. If the data are classified with respect to an attribute which can be put into two distinct categories like male-female, tall-short etc, the classification is called 'simple or dichotomous'. If the data are classified into more than two distinct categories, it is called 'mani-fold classification'.

SIMPLE CLASSIFICATION **MANIFOLD CLASSIFICATION**



of a class in an examination can be classified as :

MARKS OF 100 STUDENTS OF A CLASS

Marks	No of students
0 – 19	8
20 – 39	20
40 – 59	42
60 – 79	21
80 – 100	9
Total	100

In the above classification, the marks is the 'variable' and the number of students in each class is the 'frequency'. Such a classification is called a 'grouped frequency distribution'.

In the quantitative classifications or classifications by variables, the variables refer to the characteristics which vary. It can be either 'continuous' or 'discrete' (i.e. discontinuous). The variables those can take all possible real values within a specified interval are called 'continuous variables'. The heights of a group of individuals, their weight, their ages etc are all examples of continuous variables. For a continuous variable, the data are obtained by measurements and not by counting. The number of values that a continuous variable can take in an interval, however small it may be, is infinite. A 'discrete variable' is one that takes on some specified values (not all possible values) in an interval. For example, number of defective items in a lot, number of students remaining absent on various days during a month, number of heads obtained by throwing ten symmetrical coins, number of road accidents during a year etc. are discrete variables. Usually, discrete

A frequency distribution may be 'ungrouped' or 'grouped'. Arrangement of raw data in ascending or descending order of magnitude is called 'arraying' of data. Arraying of data is an improved version of the raw data.

3B.2.1 Ungrouped Frequency Distribution

A classification showing the different values of a variable and their respective number of occurrences (called frequencies) is termed as 'ungrouped frequency distribution' or 'discrete frequency distribution'. A frequency distribution is a better method of representation than arraying of data. Here, the number of times a particular value of the variable occurs in the data is counted which is facilitated through 'Tally marks or Tally Bars'. A tally mark is a vertical bar (or stroke) put against the value of the variable whenever it occurs. In some cases, horizontal bars are used. The following example explains the formation of an ungrouped frequency distribution.

Example. 3B.1 The marks of 60 students in a class in Statistics are given below. Form an ungrouped frequency distribution table.

Marks : 15 55 18 25 56 39 26 18 32 15 25 25 22 25 46

46 25 2 36 35 35 68 35 32 38 56 32 22 46 48

10 75 42 36 24 64 39 35 64 42 45 40 56 48 18

78 42 54 47 54 50 20 63 45 35 26 54 58 35 68

10	I	1
15	II	2
18	III	3
20	I	1
22	II	2
24	I	1
25	卌	5
26	II	2
32	III	3
35	卌 I	6
36	II	2
38	I	1
39	II	2
40	I	1
42	III	3
46	III	3
47	I	1
48	II	2
50	I	1
54	III	3
55	I	1
56	III	3
58	I	1
63	I	1
64	II	2
68	II	2
75	I	1
78	I	1
Total		60

Explanation :

In the above table No.3B.1, the first column (1) consists of all the possible values of the variable i.e. marks. The second column (2) consists of tally marks which have been given in the form of vertical strokes, one stroke for each mark counted in the data. When a particular mark is repeated for the fifth time, the fifth stroke has been marked by a cross tally mark (\) running diagonally through the previous four vertical parallel strokes. This technique facilitates counting the tally marks at the end. The total of the tally marks against each value is its 'frequency', given in the third column (3). Such a presentation of raw data is called a 'simple frequency distribution' or an 'ungrouped frequency distribution'. Here the identity of the values in the data remain unchanged but the order in which they occur remain unknown.

from the tally marks. Such a presentation of data is called 'grouped frequency distribution'. In this type of presentation of data the identity of the original figures and the order of their occurrence in the data are lost and the frequency of each class is assumed to be uniformly spread over the entire class interval.

Example 3B.2 Consider the data given in example 3B.1. This data can be put in a more condensed form in a grouped frequency distribution as follows :

Table No. 3B.2

MARKS OF 60 STUDENTS IN STATISTICS

Marks	Tally marks	Frequency
0 – 9	I	1
10 – 19	卌 I	6
20 – 29	卌 卌 I	11
30 – 39	卌 卌 IIII	14
40 – 49	卌 卌 II	12
50 – 59	卌 IIII	9
60 – 69	卌	5
70 – 79	II	2
Total		**60**

integral as well as fractional, in an interval, the formation of the frequency distributior would be such as is given in the following example.

Example 3B.3 The heights of 60 students varying between 155.3 cms and 164.8 cms have been put in the following frequency distribution.

Table No.3B.3

HEIGHTS OF 60 STUDENTS

Height in cms	Tally marks	Frequency
155 – 157	卌 III	8
157 – 159	卌 卌 IIII	14
159 – 161	卌 卌 卌 卌 I	21
161 – 163	卌 卌 I	11
163 – 165	卌 I	6
Total		60

This is called a frequency distribution with 'exclusive class intervals' or 'overlapping class intervals'. In this case, one of the limits of the class intervals (either lower or upper) is excluded. Usually, the upper limit is excluded. For example, in the above table No. 3B.3, all the heights measuring from 159 cms to a height less than 161 cms in the data would be included in the class interval marked 159 – 161. But it is inconvenient to make such tables because a person 160.8 cms tall is usually recorded to have a height of 161cm and the classification is vitiated.

3B.3 GUIDELINES FOR GROUPED FREQUENCY DISTRIBUTION

The following guidelines must be observed for a good frequency distribution of data.

The above guidelines are explained below in details.

(i) Definition of the classes : The classes should be defined clearly without ambiguity. They should be exhaustive and mutually exclusive. There should not be any confusion for the compiler as to whether a particular figure of the data would be included in one class or the other. In other words, the classes should be so defined that a particular unit of the data would belong to one class only.

(ii) Determination of the number of classes -:

There is no hard and fast rule as to what should be the total number of classes in a grouped frequency distribution. Ordinarily the number of classes should lie between 6 to 25 and ideally, between 8 and 15. The number of classes depends on the total number of observations. If the total number of observations in the data is large, the number of classes may be large because in such cases all the class intervals would contain a fairly large frequency. If the total number of observations in the data is small, the number of classes should be small or else there would be some classes containing no frequency and some others containing very low frequency.

Further, if the number of class intervals is small, the spread of the class intervals become large and too many units would be crowded in a single class. This might obscure some of the important features and characteristics of the data and might lead to loss of information and loss of accuracy.

If the number of class intervals is large, all the characteristics contained in the data would be retained, but too small frequencies would be contained in some classes and hence the purpose of condensation of data through classification would be lost. Handling of such frequency distributions would become tedious, time consuming and inconvenient· without proportionate gain in accuracy. So, from practical point of view, the number of class intervals for a frequency distribution should not be less than 6 nor should be greater than 25.

For example, if N = 100, then

$$\begin{aligned} K &= 1 + 3.322 \log_{10} 100 \\ &= 1 + 3.322 \times 2 \\ &= 1 + 6.644 \\ &= 7.644 \\ &= 8 \quad \text{(as the number of class intervals should be an integer)} \end{aligned}$$

Range: The differance between the largest and the smallest values of the data is called the range. Thus,

$$\text{Range} = L - S$$

where L is the largest value and S is the smallest value of the data.

Mid Point of the class (Class Mark) : In a grouped frequency distribution, it is assumed that the frequency of a particular class interval condenses at the centre of the class interval, though such a value may not actually exist in the data. The middle point of a class interval is considered as the representative of all the values belonging to that class interval and is called the class mark. For a class, it is computed as

$$\text{Mid point (or classmark)} = \frac{1}{2} \text{(Lower limit + Upper limit)}$$

$$= \text{Lower limit} + \frac{1}{2} \text{(Upper limit - Lower limit)}$$

(iii) Choice of class limits and class boundaries:

We know that in a grouped frequency distribution, each class is bounded by two numbers, called class limits or the class boundaries. The smaller number is called the lower limit and the larger the upper limit of the class. These two limits indicate the smallest and the largest figures in the data to be included in the class (and the figures which lie outside of these two limits can not be included in the class). In case of frequency distributions with exclusive classification, the class limits and the class boundaries are between the upper limit of a class and the lower limit of the next higher class and then adding this value to the upper limit of the class and subtracting from the lower limit of the next higher class as follows :

Inclusive classification	Exclusive classification
with class limits	with class boundaries
20-29	19.5-29.5
30-39	29.5-39.5
40-49	39.5-49.5
50-59	49.5-59.5

Finally, the lower boundary of the smallest class interval and the upper boundary of the largest class interval are determined by manipulation in par with the upper and lower boundaries respectively of other class intervals.

(iv) Determination of the magnitude of the class interval :

The difference between the two boundaries of a class is called the magnitude of the class interval. The magnitude of the class intervals should preferably be equal and can take a value like 2, 5, 10, 100, 500 etc. or a multiple of 5.

The magnitude or width of the class depends on a number of factors like the range of the data, the total frequency, the details required in the enquiry, the number of class intervals, degree of accuracy required and computational ease in further processings. An appropriate value of the magnitude of class interval can be obtained by the following formula due to Sturges.

$$\text{Magnitude of the class interval, } i = \frac{\text{Range}}{\text{No of class itntervals}} = \frac{L - S}{1 + 3.322 \log_{10} N}$$

where L and S are respectively the largest and the smallest observations and N, the total number of observations of the data.

In a frequency distribution, the frequency of a class interval indicates the number of units of the data included in that class interval. It is determined by counting either using tally marks or by using mechanical devices. How counting is facilitated by the use of tally marks has been discussed in 3B.2.1 . Counting is also possible through devices like punched cards or by use of computers.

3B.4. RELATIVE FREQUENCY, PERCENTAGE FREQUENCY AND FREQUENCY DENSITY

For special purposes, some times, relative frequencies, percentage frequencies and frequency densities of a frequency distribution are calculated.

$$\text{Relative frequency of a class} = \frac{\text{Frequency of the class}}{\text{Total frequency}}$$

$$\text{i.e. relative frequency of the ith class} = \frac{f_i}{N},$$

where f_i is the frequency of the i th class and N is the total frequency of the frequency distribution.

$$\text{Percentage frequency of the i th class} = \frac{f_i}{N} \times 100$$

where f_i is the frequency of the i th class and N is the total frequency of the frequency distribution.

Relative frequencies and percentage frequencies are used for the purpose of comparison between the classes of the same freqeuncy distribution or between the classes of two or more frequency distributions.

When the frequency distribution consists of unequal class intervals, for the purpose of comparison of concentration of frequencies in different classes, frequency densities of where f_i is the frequency and w_i is the width of the i th class.

For sake of convenience in using the frequency densities, a constant number, k is multiplied so that the frequency densities are expressed as integers. Thus,

$$\text{Frequency Density of the ith class} = \frac{f_i}{w_i} \times k$$

k may be a number like 100, 1000, etc or any other number so that $\frac{f_i}{w_i} \times k$ is an integer.

3B.5 TYPES OF FREQUENCY DISTRIBUTIONS

We give below the specimens of frequency distributions to illustrate the various ways of presentation of data.

(i) Table No. 3B.4

HEIGHTS OF 100 STUDENTS IN A CLASS

Heights in cms	No of students
120 and above but below 125	10
125 and above but below 130	18
130 and above but below 135	25
135 and above but below 140	22
140 and above but below 145	16
145 and above but below 150	9
Total	100

Weekly wages in Rs.	80 -	85 -	90 -	95 -	100 -	105 -	110 -	115 - 120	Total
No of workers	12	23	35	48	36	21	15	10	200

Here the class intervals are of equal width and we read the class intervals as 80 and above but less than 85, 85 and above but less than 90 etc. In this case the class intervals are non-overlapping.

(iii) Table No. 3B.6

ANNUAL INCOME OF 600 EMPLOYEES OF AN INDUSTRY

Annual Income in Rs.'000	No. of employees
Above 20 but not exceeding 30	15
Above 30 but not exceeding 40	80
Above 40 but not exceeding 50	120
Above 50 but not exceeding 60	180
Above 60 but not exceeding 70	110
Above 70 but not exceeding 80	70
Above 80 but not exceeding 90	25
Total	600

Here the first class begins with a number above Rs.20,000 and ends with a number Rs.30,000 and implies that the lower limit is not included while the upper limit of the class interval is included. Implications for other classes would be similar. The limits of the class intervals are thus non-overlapping.

This is a frequency distribution with exclusive classification and of unequal class intervals. The magnitude of the class intervals are Rs.50, 50, 100, 150, 200, 150 lakhs respectively. Here the first class interval is read as Rs. 100 lakhs and above but less than Rs.150 lakhs. The other class intervals are read accordingly. Here the upper limits of the classes are excluded from the corresponding class intervals while the lower limits are included. Unequal class intervals are preferred when there exists greater variation in the data i.e. when there is sharp rise or fall in the frequency over a small interval.

Example 3B.4. The marks of 300 students in an examination in Mathematics are given in the following frequency distribution.

Marks in Math	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	Total
No. of Students	6	15	32	145	53	28	14	7	300

Find

(i) the class boundaries and the mid-point of the fourth class.

(ii) magnitude of this class interval.

(iii) percentage of students having 60 or more marks.

Solution :

(i) Since the frequency distribution is given in inclusive classes, the class boundaries and the class limits are not the same. Now we find that the fourth class interval is 50-59. Then we look at the preceding and suceeding class intervals which are 40-49 and 60-69 respectively. Thus, the three consecutive class intervals are

40-49 50-59 60-69

The gaps between the preceding and succeeding class intervals are 1 each i.e. 50-49=1

Mid point of the fourth class = $\frac{59+50}{2} = 54.5$ marks.

(ii) The magnitude of the class interval = Upper boundary – Lower boundary

$= 59.5 - 49.5 = 10$ marks.

(iii) As the marks are only integral values, we assume that there is no fractional value between 59 and 60. So, the number of students having 60 or more marks is the total number of students who belong to the class intervals from 60-69 and onwards. Thus, the number of students having 60 or more marks = 53 + 28 + 14 + 7=102

Total number of students = 300

Percentage of students having 60 or more marks

$$\frac{\text{No. of students having 60 or more marks}}{\text{Total number of students}} \times 100$$

$$= \frac{102}{300} \times 100$$

$$= 34$$

Example 3B.5. The following frequency distribution presents the marks of 108 students of a class in a special test.

Marks	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	Total
No of Students	4	7	10	20	22	18	12	10	5	108

It was decided to group the students into three different categories:

(a) those whose marks were less than 55.

(i) How many students would be given special coaching?

(ii) Howmany students would belong to category (b)?

(iii) Howmuch money would be spent on prizes?

Solutions :

For solution of such problems where interpolation is necessary, the data, if given in the foam of inclusive classification, are to be expressed to exclusive forms so that the assumptions are tenable. After conversion, the frequency distribution becomes :

Marks	9.5-19.5	19.5-29.5	29.5-39.5	39.5-49.5	49.5-59.5	59.5-69.5	69.5-79.5	79.5-89.5	89.5-99.5	Total
No of Student	4	7	10	20	22	18	12	10	5	108

The following assumptions are in order.

1. The class frequency of a class is uniformly spread over the entire class interval of the corresponding class.

2. The frequency of each class is condensed at the mid point of the class.

(i) The number of students in the category (a) is found by adding all the frequencies as follows :

The number of students who are to be given special coaching

$$= 4+7+10+20+\frac{55-49.5}{59.5-49.5}\times 22$$

$$= 41+\frac{5.5}{10}\times 22$$

$$= 41 + 12.1$$

$$= 53.1 = 53$$

$= 28.5 = 29$

(iii) Total money prize

= No of students belonging to category (c) x Rs500.

$$= \left[\frac{79.5-77}{79.5-69.5}\times 12+10+5\right]\times 500$$

$= 18 \times 500$

= Rs. 9000.

3B.5(a) CUMULATIVE FREQUENCY DISTRIBUTION

In a frequency distribution, the frequency of the values of the variable or of the class intervals indicate how frequently the different values of a group occur in the data. For the purpose of graphical representation or for finding positional averages or for some other purposes, it may be necessary to count the number of units of the data having values less than or greater than a given figure. To handle such situations, accumulated frequencies for various values are counted by adding all those frequencies up to the desired value either from the top or from the bottom of the frequency distribution. These accumulated frequencies are called 'cumulative frequencies'. The corresponding cumulative frequency is shown in a separate column against each value or class interval as the case may be. Such a display is called 'cumulative frequency distribution'. Cumulative frequency distributions can be of two types viz. (i) less than cumulative frequency distribution and (ii) more than cumulative frequency distribution.

Corresponding to a specified value of an ungrouped frequency distribution, the less than cumulative frequency is the number of units (values or measurements) whose values are less than or equal to that value and the more than cumulative frequency is the number of units whose values are greater than or equal to that value. For a grouped frequency distribution the cumulative frequencies are shown against the class boundary points. Less than cumulative frequency of a class shows the total number of units of the more than cumulative frequency of a class interval is the sum of the frequencies of all the class intervals which are equal to or greater than that class interval.

LESS THAN CUMULATIVE FREQUENCY DISTRIBUTION

Marks	Class boundaries of the marks	Frequency	Less than cum frequency (Less than upper boundary)
20-29	19.5-29.5	8	8
30-39	29.5-39-5	15	8+15=23
40-49	39.5-49.5	20	23+20=43
50-59	49.5-59.5	35	43+35=78
60-69	59.5-69.5	17	78+17=95
70-79	69.5-79.5	12	95+12=107
80-89	79.5-89.5	8	107+8=115
90-99	89.5-99.5	5	115+5=120
Total		120	

MORE THAN CUMULATIVE FREQUENCY DISTRIBUTION

Wts in kgm	class boundaries	(No of persons) Frequency	More than cum. frequency. (More than the lower boundary)
40-44	39.5-44.5	12	108+12=120
45-49	44.5-49.5	23	85+23=108
50-54	49.5-54.5	38	47+38=85
55-59	54.5-59.5	24	23+24=47
60-64	59.5-64.5	14	9+14=23
65-69	64.5-69.5	9	9
Total		120	

(8+15) is 23, and so on.

For more than cumulative frequency distribution, we have considered an inclusive classification of data (exclusive classification can be used) and so in this case also, the calss boundaries have been determined as before. Here the cumulative frequencies have been calculated from the bottom of the frequency distribution. For example, for the class boundary 64.5-69.5, the frequency is 9 and the c.f. is 9. For the class 59.5-64.5, the frequency is 14 and the c.f. (9+14) is 23 and so on. The cumulative frequency for the class 39.5-44.5 is 120 which shows that all the 120 persons have weights 39.5 kg or more.

Uses : Cumulative frequencies are used to determine the number of units (or the values of the variable) in the data which are less than or more than a specified value and also to find the number of units those lie between two specified values. For computation of partition values like median, quartiles, deciles etc. cumulative frequencies are used.

3B.6 BIVARIATE FREQUENCY DISTRIBUTION

We have so far discussed about data relating to one variable only where the variables like height, weight, income, production, marks etc. were considered. The frequency distributions of such data are called 'Univariate' frequency distributions. When we consider populations with two variables e.g. height and weight, income and expenditure, marks in Statistics and in Mathematics etc., such populations are called 'Bivariate' populations and the data collected from those populations are called bivariate data. Each unit of the data consists of a pair of measurements or values and is denoted by (x, y) where x refers to the value or measurement of one variable and y, the other. Thus, the pair (x_i, y_i) represents the pair of values for the i th unit of the data. When the bivariate data is presented in the form of a frequency distribution, it is called a 'bivariate frequency distribution'. Bivariate frequency distributions are presented in the form of two-way tables.

illustrate how to form a bivariate frequency distribution.

Example 3B.6. The ages of 20 couples are given below. Represent the data in a bivariate frequency distribution.

Ages in years : (28,23), (37,30), (42,40), (25,26), (29,25), (47,41), (37,35), (35,25), (23,21), (41,38), (27,24), (39,34), (23,20), (33,31), (36,29), (32,35), (22,23), (29,27), (38,34), (48,47)

Solution :

For each pair of value, the first number refers to the age of the husband and the second, the age of the wife.

Table No.3B.9

BIVARIATE FREQUENCY DISTRIBUTION OF THE AGES

OF 20 COUPLES

Age of husband	Age of wife						
	19.5-24.5	24.5-29.5	29.5-34.5	34.5-39.5	39.5-44.5	44.5-49.5	Total
19.5-24.5	III ③	–	–	–	–	–	3
24.5-29.5	II ②	III ③	–	–	–	–	5
29.5-34.5	–	–	I ①	I ①	–	–	2
34.5-39.5	–	II ②	III ③	I ①	–	–	6
39.5-44.5	–	–	–	I ①	I ①	–	2
44.5-49.5	–	–	–	–	I ①	I ①	2
Total	5	5	4	3	2	1	20

Explanation : In the table above, the figures in the circles of the respective cells are the frequencies. The class intervals are exclusive classification in which the upper limits of the classes are excluded while the lower limits are included for the purpose of counting. Tally marks are put as follows :

and written in circles at the right hand top corner of the respective cells. Note that the totals written in the extreme right column are the row totals which indicate the number of husbands in the corresponding age group and the totals appearing in the bottom row are the column totals which indicate the number of wives of the corresponding age group. These row and column totals are called 'marginal totals'. The sum of the marginal totals of the rows and the sum of the marginal totals of the columns are equal and in turn are equal to the total number of pairs of data.

3B.7. TABULATION OF DATA

3B.7.1 Meaning and definition

Tabulation refers to systematic presentation of data in the form of tables consisting of rows and columns. It helps in condensation of data, brings clarity and facilitates comparison.

A. M. Turtle defines tabulation as "The logical listing of related quantitative data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, phrases and statements in the form of titles, headings and explanatory notes to make clear the full meaning, context and the origin of the data."

This definition states the structure of statistical tables and suggests that tabulation highlights the basic characteristics of the information contained in the data through its orderly and systematic arrangement.

3B.7.2 Difference between Classification and Tabulation.

Classification is the first step before tabulation. So, both classification and tabulation go together. Through classification similar units are grouped together. Only after classification, tabulation is taken up to display data under different columns and rows so that it can be easily understood and compared.

which faciliate comparison between different parts of the table.

3. **To facilitate statistical analysis:** After classification and tabulation, statistical data become fit for analysis and interpretation. Various descriptive statistics like mean, median, standard deviation etc. can be calculated easily from systematically tabulated data.

4. **To provide identity of the data:** The table number, the rows and columns of the table can be referred to in future studies.

5. **To depict the trend and pattern of the data:** Through tabulation, the patterns within the figures can be revealed which can not be observed from the raw data or its description in words.

3B.7.4 Essential parts of a Table

Depending upon the nature of the data and the purpose of the study, a table is divided into different parts. We state below the essential parts for a general purpose table.

1. **Table Number :** For the purpose of easy identification, each table should be assigned with a specific separate number. This number is usually placed at the centre of the top, above the title or at the left hand side top corner or at the bottom of the table in the centre.

2. **Title of the Table :** Every table must have a suitable title indicating the contents of the table. The title should be clear and precise and should be written prominently, preferably in bold letters and must be given at the top of the table at the centre.

3. **Captions :** Caption refers to column headings. Each column heading may have a number of sub headings. The captions should be clearly defined and placed at the middle of the column top. In case, various columns are expressed in different units, the same should be indicated along with the captions.

6. **Head Note or Prefatory Note :** Head note refers to the data contained in the major part of the table. It is placed below the title of the table in the form of brief explanatory statement but in brackets.

7. **Foot Note :** Footnotes are given below the table. These are in the form of statements meant to clarify any thing that has not been clear from the heading, title, captions, stubs etc.

8. **Source :** Source of the data must be mentioned at the bottom of the table.

SPECIMEN OF A TABLE

Table No ....

Title ....

Head note .....

Stub Headings	← Captions →				
	Column Heading	Column Heading	Column Heading	Column Heading	Total
Stub Entries ↓	B	O	D	Y	
Total					

Foot Note ....

Source ....

1. The table should be precise and easy to understand. It may be such that one does not need to read the foot notes or explanations to properly understand the table.
2. The table should suit to the size of the paper. The width of the columns should be so determined that they can be accommodated in the available space of the paper.
3. Very large data should not be crowded in a single table. Such data should be presented in separate tables.
4. There should be a proper title of the table indicating what exactly it represents.
5. The main headings should be few. However, number of sub-headings may be large.
6. Headings of columns (captions) and rows (stubs) should be self explanatory i.e. what the headings represent should be mentioned.
7. The columns where data are to be compared should be placed side by side. Percentages, totals and averages must be kept close to the data.
8. To reduce unnecessary details, the figures should be approximated before tabulation.
9. Under each heading, the unit of measurement, if any, must be indicated.
10. Row totals should be kept at the extreme right column and column totals at the bottom of the table. In some cases, the row totals may be placed at the first column and the column totals at the top of the table.
11. If some figures are to be emphasized, they should be in distinctive type i.e. may be written in bold numerals or kept in boxes or encircled or put between thick lines.
12. In case, percentages are to be kept side by side with original figures, they should be in a separate type of numerals i.e. italics.
13. To present portion of collected facts which remain unrepresented in the table, a miscellaneous category should be created.

be given explicitly Ditto marks should not be used in the table.

16. Abbreviations should be avoided in the title and subtitles.

17. Source of the data must be mentioned.

It may be difficult to follow all these rules while preparing a table, but their purpose should always be kept in mind.

3B.7.6 Types of Tables

All tables may be put under two categories viz (i) Simple and complex table and (ii) General purpose and special purpose table.

(i) Simple and complex table

Depending upon the number of characteristics displayed, a table is called simple table or complex table. In a simple table, only one characteristic is exhibited. Such tables are also called 'One - way' table. A table showing the distribution of marks of a group of children in a class is a simple or one - way table. On the other hand, a table showing two or more characteristics simultaneously is called a complex table. A complex table may be 'two - way' or 'three - way' and so on. When two characteristics are exhibited simultaneously in a table, it is called a two - way table. For example, a table showing the income and age of 500 persons in a locality is a two - way table. Like wise, for more than two characteristics, higher order tables can be formed.

25 - < 35	20
35 - < 45	30
45 - < 55	20
55 and above	10
Total	100

Two – way table

NO. OF EMPLOYEES IN AN ORGANIZATION ACCORDING TO AGE AND SEX.

Age in years	Employees		
	Male	Female	Total
Below 25	8	4	12
25 - < 35	15	13	28
35 - < 45	20	10	30
45 - < 55	17	3	20
55 and above	5	5	10
Total	65	35	100

(ii) General and special purpose table

General purpose tables are also called reference tables or repository tables. These are meant for general use or reference only. These tables contain information which are not meant for a particular purpose. For example, the detailed tables containing the population census of India are tables of this category.

can be obtained from general purpose tables in the form of ratios, percentages and other measures.

Example. 3B.7

In 1995, out of a total of 2000 students in a college, 1400 were for graduation and the rest for post graduation. Out of 1400 graduate students, 100 were girls; however in all there were 600 girls in the college.

In 2000, the number of graduate students increased to 1700 out of which 250 were girls, but the number of post graduate students fell to 500, out of which only 50 were boys.

In 2005, out of 800 girl students, 650 were for graduation, where as the total number of students was 2200. The number of boys and girls in graduate classes were equal.

Represent the above information in a tabular form. Also calculate the percentage increase in the number of graduate students in the college in 2005 as compared to 1995.

(Identical to IAS Exam 1997)

Table No. 3B.10

CLASSWISE AND SEXWISE DISTRIBUTION OF STUDENTS OF A COLLEGE

Year	Classes								
	Graduate class			Post graduate class			Total		
	Boys (1)	Girls (2)	Total (3)	Boys (4)	Girls (5)	Total (6)	Boys (7)	Girls (8)	Total (9)
1995	(3) – (2) 1300	100	1400	(7)–(1) 100	(8) – (2) 500	(9) – (3) 600	(9) – (8) 1400	600	2000
2000	(3) – (2) 1450	250	1700	50	(6) – (4) 450	500	(9) – (8) 1500	700	2200
2005	650	650	(1) + (2) 1300	(7) – (1) 750	(6) – (4) 150	(9) – (3) 900	(9) – (8) 1400	800	2200

Column (2) = (7) – (1) = 1400 – 1300 = 100

Similarly for the years 2000 and 2005 the figures in the vacant spaces have been calculated as indicated in the table.

Percent increase in number of graduate students in 2005 as compared to 1995

$$= \frac{\text{No. of graduate students in 2005} - \text{No of graduate students in 1995}}{\text{No. of graduate students in 1995}} \times 100$$

$$= \frac{1300 - 1400}{1400} \times 100$$

$$= -\frac{100}{14}$$

$$= -7.14\ \%$$

So, there has been 7.14% decrease in the number of graduate students in the year 2005 as compared to the year 1995 in the college.

3B.8 MACHINE TABULATION

Tabulation of data can be done either using hand operated or electrically operated machines. Method of 'needle sorting' is performed by hand operated machines. In this method, a large number of units of the data can be sorted under any number of headings and sub headings. Cards of convenient shape and size with a series of holes, each hole representing a value, are used. After stacking the cards, a needle is made to pass through a particular hole representing a particular value of the variable. All those cards through which the needle passes are then separated and counted. The number of cards thus counted represent the frequency of the value of the variable (or of the class interval). In this way, the frequencies of all the class intervals are counted by repeating the process.

accurate and automatic.

Advantages : The advantages of mechanical tabulation are :

(i) Data can be tabulated in a short time.

(ii) Greater accuracy can be ensured.

(iii) Cost can be reduced considerably.

(iv) Large scale data can be handled

(v) Monotonous and un-interesting work of tabulation of data is transferred to machines there by involving less human labour and minimizing error.

3B.9 DIAGRAMS, GRAPHS AND CHARTS

We have already discussed in the earlier sections, the techniques of presenting statistical data through classification and tabulation. Classification and tabulation reduce the complexity of data and present those systematically in an intelligible form. Diagrams, graphs and charts are other important methods of presentation of statistical data in a convincing and more appealing form which can be easily understood by common people. The speciality of diagrams and graphs is that they present the numerical figures in the shape of attractive and appealing pictures and charts which attract the attention of common man.

3B.9.1. Advantages of Diagrams and Graphs.

Some of the advantages of diagrams and graphs are :

(i) **They are visual aids which give a bird's eye view of the numerical facts and present them in simple and readily comprehensible form.**

(ii) **They create deeper impressions on the minds of the readers.** The impression created by a diagram or picture is supposed to last long in the watching the numbers may be revealed through diagrams and graphs.

(iv) **They save time.** To draw a conclusion by comparing a set of numbers one needs time. But such conclusion can be drawn at once by looking at a diagram.

(v) **They reveal trend.** Presence of trend in the data may not be identified from the numericals of the data. Such identification may often be difficult or even impossible. But graphs reveal the trends and exhibit the manner in which the trends change.

3B.9.2 Limitations

(i) **They give only approximate picture of the data.** Diagrams provide only approximate picture and so are used only when the purpose is to explain the significance of some statistical facts to common people. Hence these are not of much importance for a person doing exhaustive study of the numbers.

(ii) **They can be used only for comparative study.** A single diagram has no significance for comparison. Comparison is possible if there are two or more diagrams.

(iii) **They can be misused easily.** Presentation of data by wrong and inappropriate diagrams can lead to fallacious conclusion.

3B.9.3 Difference between Diagrams and Graphs.

(i) Graphs are drawn on graph papers only. They help in studying the mathematical relationship between the variables. But diagrams are usually drawn on plain papers and are useful for comparison.

(ii) Graphs are precise and more accurate than diagrams. They can be used for obtaining intermediate values and projecting future values. Diagrams furnish only approximate information.

title conveying the main theme which is desired to be portrayed through it. Usually the title is placed at the top though sometimes it may be put at the bottom. If necessary, in order to clarify some points relating to the diagram, foot notes may be given at the lefthand side bottom. Besides, the source of the data for which the diagram is drawn should be mentioned.

(ii) **Size.** The size of the diagram depends on the quantum of data and should be such that all the characteristics desired to be shown are properly emphasized and is understood by a mere look.

(iii) **Proportion between width and height.** Appropriate proportion between the width and height should be maintained, although there is no fixed rule for this proportions.

(iv) **Scale.** The scale to be used in drawing a diagram depends on the magnitude of the data and the size of the paper. The scale chosen should be an even number or preferably a multiple of 5, 10, 20, 25 etc. It should specify the size of the unit of the data it represents.

(v) **Look.** The diagram should be neatly drawn and should look attractive to the eye. The diagram which attracts the attention of the common man is called a good diagram.

(vi) **Index.** An index explaining the different types of lines, shades or colours used should be given for easy understanding.

(vii) **Simplicity.** Diagrams should be simple. Too much information should not be crowded in a single diagram. For data with several items of information, more than one diagram should be used, each portraying some specific characteristic.

(ii) Two dimensional diagram

(iii) Three dimensional diagram

(iv) Pictogram

(v) Cartogram

3B.10.3 One Dimensional Diagram

One dimensional diagrams usually refer to 'Bar Diagrams'. The following types of bar diagrams are frequently used in presenting data.

(a) Simple Bar Diagram

(b) Multiple Bar Diagram

(c) Subdivided Bar Diagram or Component Bar Diagram

(d) Percentage Bar Diagram.

(e) Pie-Diagram

3B.10.3(a) Simple Bar Diagram

Simple bar diagrams consist of a number of equispaced vertical or horizontal rectangles. Each rectangle represents the magnitude of the value of the variable. They represent data of discrete variables relating to different places, different time periods etc. i.e. data relating to population of a place in different years, production of a manufacturer on different days, etc. For drawing a simple bar diagram, the variable such as places, periods etc are taken along the base line (x-axis) at regular intervals and the corresponding values along the ordinate (y-axis). Rectangles (or bars) of heights equal to or proportional to the corresponding values (or frequencies) of the variables as per chosen scale are erected at the points marked on the x-axis. These rectangles may be filled in with dots, dashes or colour shade.

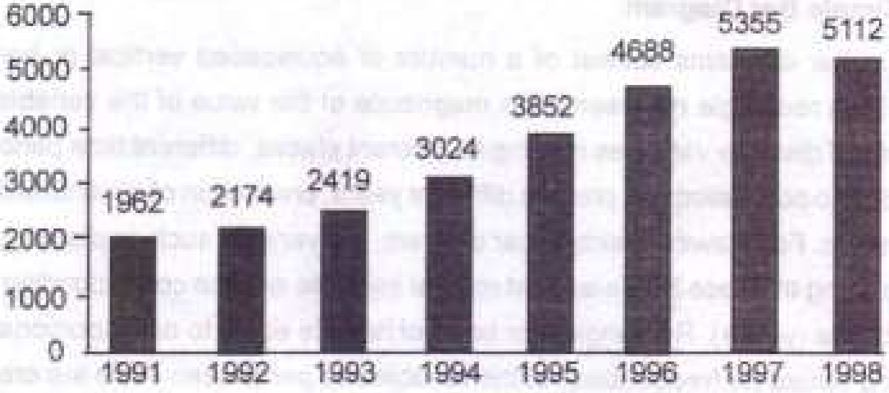
In these diagrams, only the heights of the rectangles differ from one another and the widths remain the same. The bars can be of any width but should be such that suits

1991	1962
1992	2174
1993	2419
1994	3024
1995	3852
1996	4688
1997	5355
1998	5112

Solution : The simple bar diagram is given below.

Fig - 1

TOTAL MERCHANDISE EXPORT OF INDIA DURING 1991 - 1992



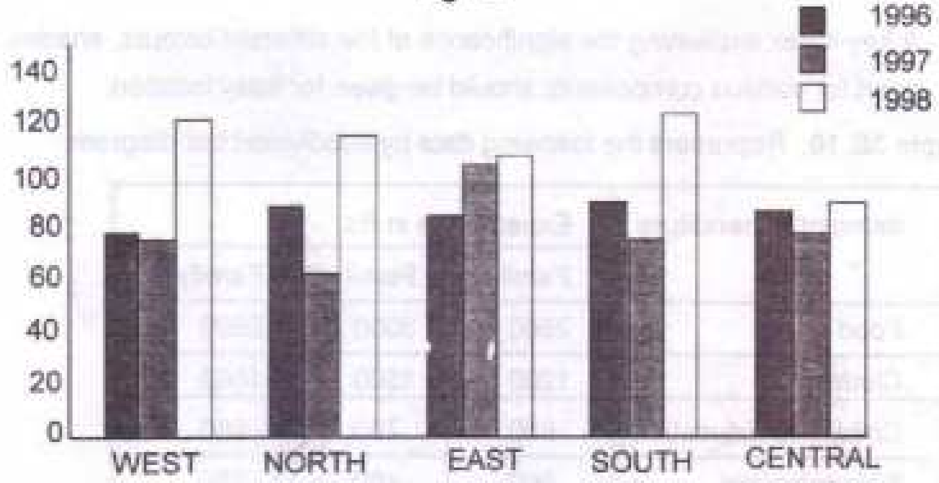
3B.10.3(b) Multiple Bar Diagram

Multiple bar diagrams are used to represent two or more sets of inter-related phenomena with respect to different places, periods or timings. They facilitate intergroup comparison. For example, to represent the sex wise population of Bhubaneswar for the to 1998 by a multiple bar diagram.

Year	Region				
	West	North	East	South	Centre
1996	78.4	88.9	83.7	89.9	86.5
1997	75.6	62.5	103.6	75.5	77.4
1998	121.2	116.5	107.6	123.9	90.3

Solution : The multiple bar diagram is given below :

Fig - 2



3B.10.3(c) Sub-divided Bar Diagram

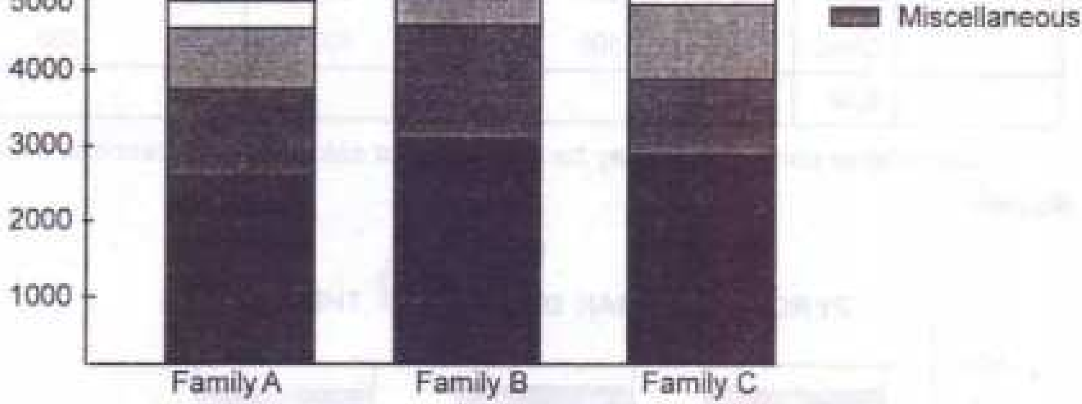
Subdivided bar diagrams are also called component bar diagram. In these diagrams, each bar which represents the magnitude of the given phenomena is further subdivided into its various components. Each component occupies a part of the bar proportional to its share in the total. These diagrams are helpful to present several items displayed is drawn. Then the whole bar is divided into segments equal to the number of components. Each segment represents a component proportional to the total of the phenomenon. Different shades, colours, dots etc are used for the identity of each component.

The following points may be kept in mind while drawing a subdivided bar diagram.

1. Order of the various components should be the same for all the bars.
2. The largest component should be shown at the base and the smallest at the top.
3. It should not be recommended if the number of components is large i.e. exceeds 10.
4. A key index explaining the significance of the different colours, shades, dots etc. used for various components should be given for easy location.

Example 3B.10. Represent the following data by subdivided bar diagram.

Items of Expenditure	Expenditure in Rs.		
	Family A	Family B	Family C
Food	2500	3000	2800
Clothing	1200	1500	1000
Children's Education	800	700	950
Entertainment	300	400	250
Miscellaneous	1200	900	1200
Total	6000	6500	6200



3B.10.3(d) Percentage Bar Diagram.

We know that in a subdivided bar diagram, the height of the bar is divided in to different segments, each segment representing a separate component. Each component occupies a part proportional to its share in the total. When the share of the components are expressed as percentages of the total and presented in the form of a bar diagram, such a diagram is termed as a percentage bar diagram. In this case, the heights of all the bars are equal to 100. These bars are usuful for portrayal of relative changes in different components to the whole of the data.

Example 3B.11. The following data are the results of the students of three colleges A. B and C at the CHSE Examination in 2005. Represent the data in percentage bar diagram

College	Result				
	I Div	II Div	III Div	Compartmental	Total
A	15	26	30	4	75
B	88	77	51	12	228
C	25	37	58	10	130

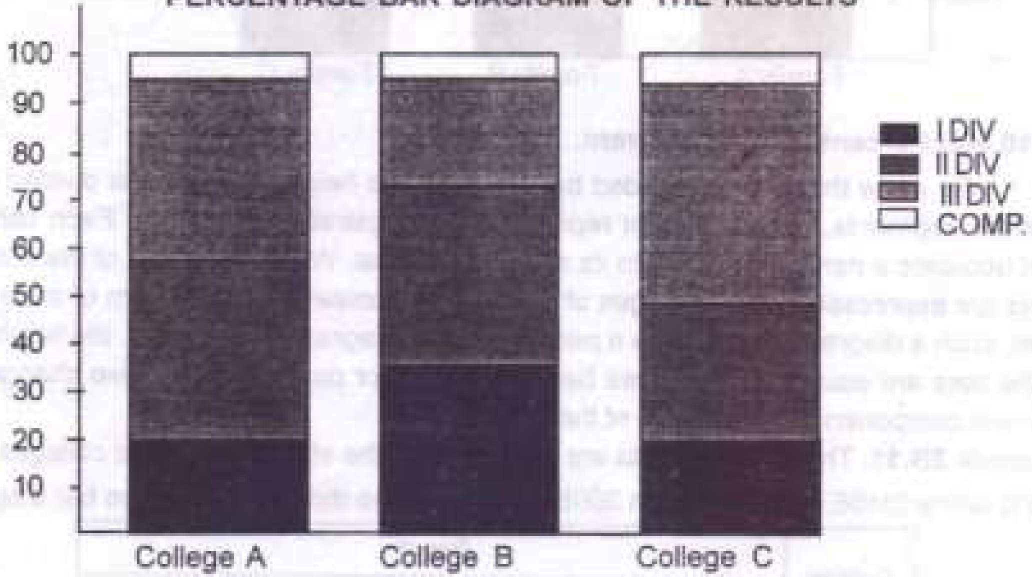
Solution : The percentages of each category of students of the colleges have been calculated in the followng table.

	III Div	40	94.7	22.4	94.8	44.5	92.3
	Comp	5.3	100	5.2	100	7.7	100
	Total	100		100		100	

Cumulative percentages may be computed for sake of convenience to draw the diagram.

Fig - 4

PERCENTAGE BAR DIAGRAM OF THE RESULTS



3B.10.4(e) Pie Diagram

Pie-diagram or Pie-chart is a very useful device for presenting the various component parts of an aggregate. It is a substitute for the subdivided bar diagram. Pie-diagrams can effectively display the comparison between the various components or between parts of the whole.

would be expressed as percentages of the whole. Then taking 100 as 360°, each of the percent of the component share would be converted to angles taking 1% as 3.6° (This is because the total angle at the centre of a circle is 360° and the total percentage of the component values is 100.). These two steps combined is same as computing the share of each component proportionately taking the total as 360°. In this way the central angles of the sectors corresponding to all the components would be determined. Then a circle with a convenient radius (that suits the size of the paper and space available) would be drawn. The area under the circle would then be divided in to different sectors using the angles computed and then each sector would be shaded with a different colour.

It is possible to determine the percentage of share of a component from a pie - diagram when the angle subtended by the sector at the centre is known. (percentage of share = Angle x $\frac{5}{18}$). For example, if the angle of the sector at the centre is 40°, then the percentage of the share of the component is 40 x $\frac{5}{18}$ =11.1% of the total.

Example. 3B.12. Draw a pie-diagram to represent the following data of the production cost of a manufacturer.

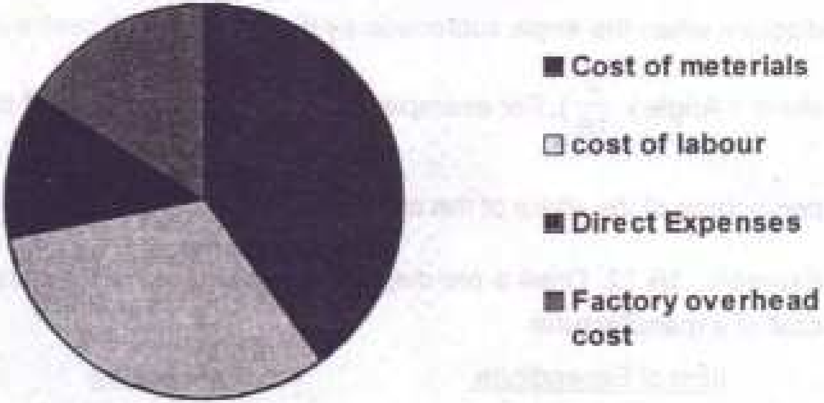
Item of Expenditure.	Expenditure
Cost of material	Rs.3,84,000
Cost of labour	Rs.3,07,200
Direct expenses of manufacturer	Rs.1,15,200
Factory overhead expenses	Rs.1,53,600
Total	Rs.9,60,000

Solution : Each item of expenditure is first converted to percentage on the basis of the total expenditure as follows :

		960000	
2.	Cost of labour	$\frac{307000}{960000} \times 100 = 32$	32 x 3.6° = 115.2°
3.	Direct Expenses	12	43.2°
4.	Factory overhead cost	16	57.6°
Total		100	360°

Fig - 5

PRODUCTION COST OF A MANUFACTURER



Limitations of Pie-Diagram

Pie-Diagrams are considered to be less effective than sub-divided bar diagrams for accurate reading and interpretation. When the data are to be presented having a number of components or the differences among the magnitudes of the components are small, such data should not be portrayed by a pie-diagram. In such cases, the relative values displayed by different sectors may lead to confusion. Although in actual practice, pie-diagrams are frequently used for different types of data, these are considered inferior to simple bar diagrams or sub-divided bar diagrams.

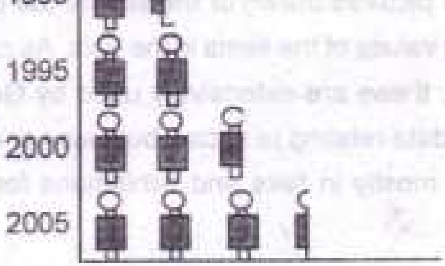
such a way that the number of pictures drawn or the size of the different pictures are in proportion to the corresponding values of the items in the data. As pictures draw immediate attention of the common man, these are extensively used by Government and private institutions for presentaiton of data relating to social, business or economic phenomena. Such diagrams are displayed mostly in fairs and exhibitions for understanding of the commmon people.

Rules for drawing pictograms

1. The picture should represent a general concept like man, woman, bus, aircraft, cows, goats, trees, books etc. and should not be the pictures of particular individuals living or dead.
2. The picture should be clear, concise and attractive.
3. Each picture should be distinct and distinguishable from every other picture.
4. The figures should suit to the size and space of the paper.
5. Changes in the values should be shown by drawing correct number of pictures.
6. The picture should be such that they convey the concept of the population under study.

Example 3B.13. Represent the followng data of student strength of a university given in quinquinnial interval.

Year	No of students in '000
1985	7
1990	10
1995	14
2000	19
2005	25

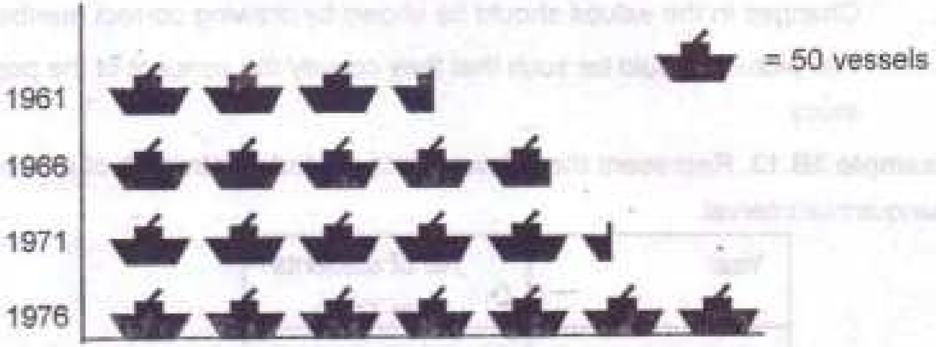


Example 3B.14. Represent the following data of number of vessels in Indian Merchant Shipping Fleet by Pictogram.

Year	1961	1966	1971	1976
No of vessel	174	231	255	350

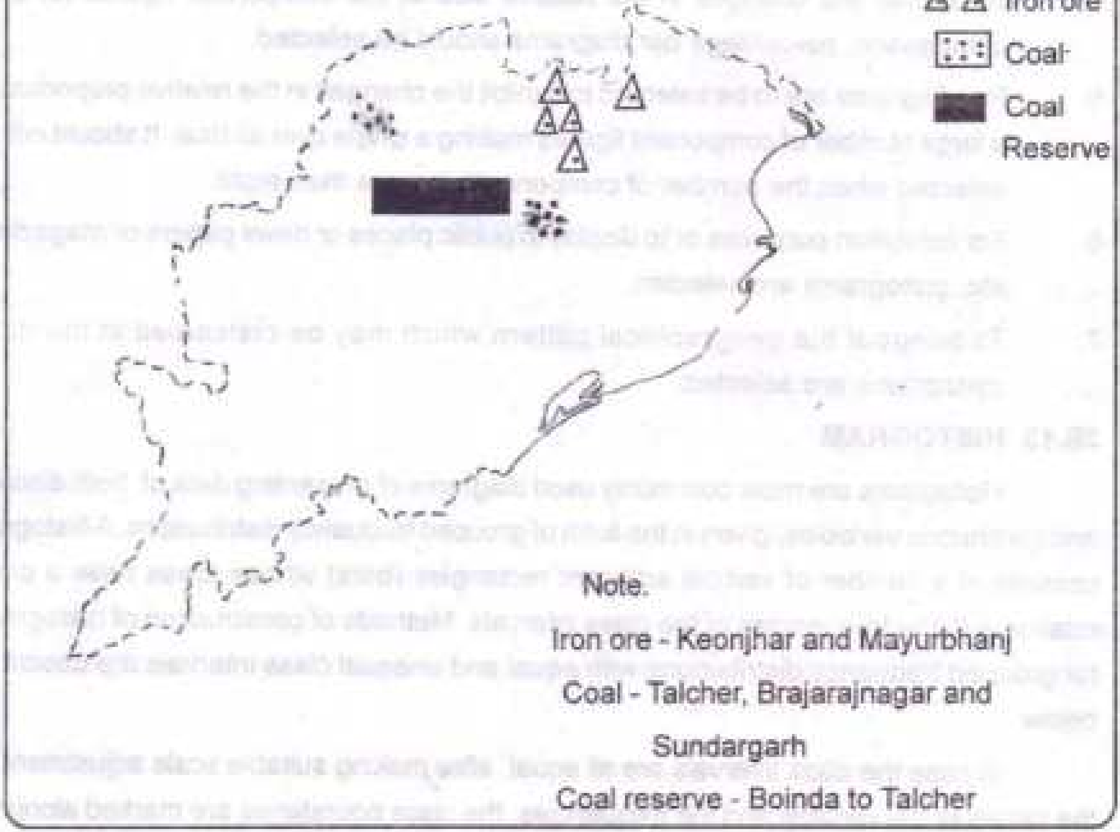
Fig - 7

VESSELS IN INDIAN MERCHANTSHIPPING FLEET



3B.11.2. Cartograms

Cartograms are statistical maps which represent quantitative information on regional or geographical basis. The distribution of rain fall in various regions of a country, the density of population in different regions or cities, the production of coal in various parts of a country etc. can be shown in these type of maps.



Note.

Iron ore - Keonjhar and Mayurbhanj

Coal - Talcher, Brajarajnagar and Sundargarh

Coal reserve - Boinda to Talcher

3B.12 CHOICE OF A SUITABLE DIAGRAM

No single diagram can be used for all purposes of presenting data. The choice of a diagram for particular data requires skill, intelligence and experience. The diagram should be selected keeping in view the nature of the data and the purpose for which it is to be displayed. A wrong choice of the diagram may lead to mis-interpretation. Some guidelines stated below may be followed while choosing a diagram.

4. To display the changes in the relative size of the component figures for their comparison, percentage bar diagrams should be selected.

5. Pie-diagrams are to be selected to exhibit the changes in the relative proportion of a large number of component figures making a single over all total. It should not be selected when the number of components is more than eight.

6. For exhibition purposes or to display in public places or news papers or magazines etc. pictograms are selected.

7. To bringout the geographical pattern which may be concealed in the data, cartograms are selected.

3B.13 HISTOGRAM

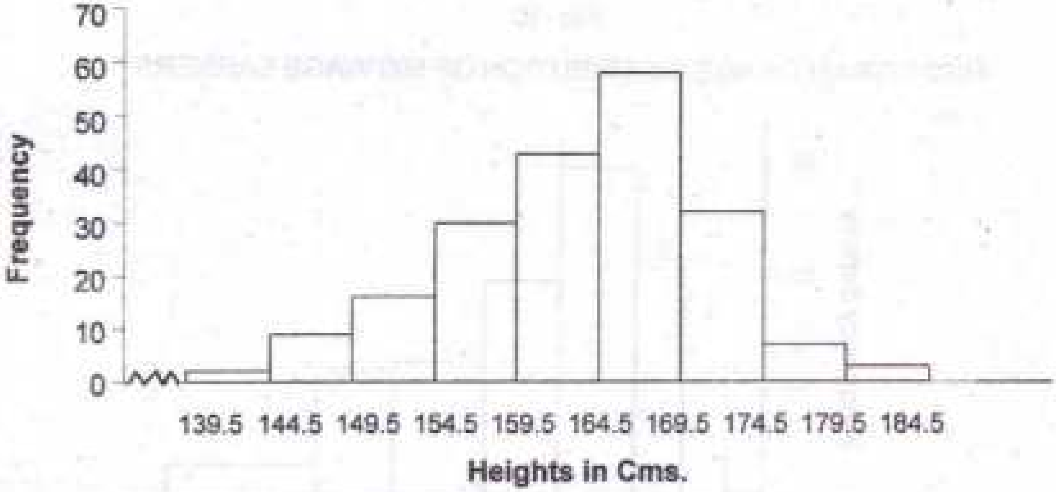
Histograms are most commonly used diagrams of presenting data of, both discrete and continuous variables, given in the form of grouped frequency distributions. A histogram consists of a number of vertical adjacent rectangles (bars) whose areas have a direct relation with the frequencies of the class intervals. Methods of construction of histograms for grouped frequency distributions with equal and unequal class intervals are described below :

In case the class intervals are all equal, after making suitable scale adjustment for the values of the variable and the frequencies, the class boundaries are marked along the horizontal line (x-axis) and the frequencies along the vertical line (y-axis). Then against each class interval, a rectangle is erected with height equal to or proportional to the corresponding frequency of the class interval. In this case, the area of the rectangle would be equal to the frequency of the corresponding class interval. For unequal class intervals, instead of taking the frequency of the class interval, the frequency density i.e. frequency per unit interval is taken as the height of the rectangle. In this case, the area of the rectangle would be proportional to the frequency of the corresponding class interval. The diagram so constructed cansists of adjacent vartical rectangles. Such a diagram is called a Histogram.

No of Students	2	9	16	30	43	58	32	7	3	200

Fig - 9

HISTOGRAM OF THE HEIGHTS OF 200 STUDENTS



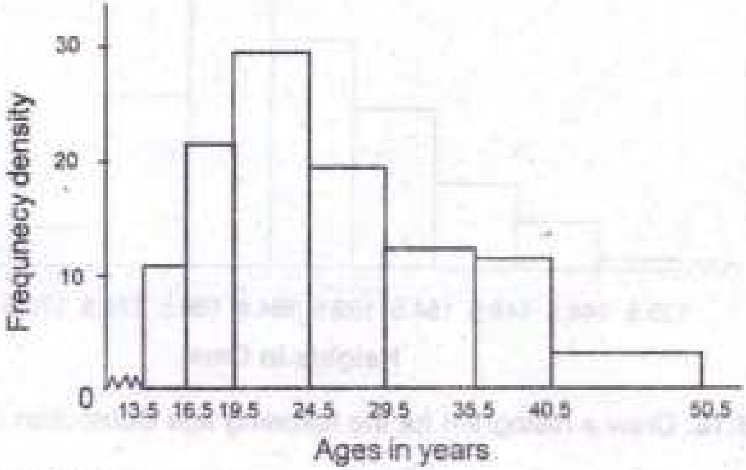
Eample 3B.16. Draw a histogram for the following age distribution of 500 workers in a factory.

Age in years	14-16	17-19	20-24	25-29	30-35	36-40	41-50	Total
No of wage earners	33	69	145	95	78	60	20	500

25-29	24.5-29.5	95	19
30-35	29.5-35.5	78	13
36-40	35.5-40.5	60	12
41-50	40.5-50.5	20	2
Total		500/.	

Fig -10

HISTOGRAM OF AGE DISTRIBUTION OF 500 WAGE EARNERS.



3B.14 Frequency polygon

Frequency polygon is a graphical presentation of a frequency distribution obtained in the form of a linegraph. It consists of an area bounded by many sides (more than four). For a grouped frequency distribution, the total area under the frequency polygon is approximately equal to the total area under the histogram. Frequency polygons are useful for comparison of two or more frequency distributions.

two extreme ends of the incomplete graph are joined by straight lines to these two extreme marked points. Such a graph having the shape of a polygon is called a frequency polygon.

For a grouped data, the frequency polygon can be drawn along with the histogram or by using the frequencies of the class intervals directly without drawing a histogram.

First a histogram is drawn for the given frequency distribution. Then the middle points of the upper edges of the rectangles for all the class intervals are located. Then the consecutive middle points, marked along the upper edges of the rectangles, are joined by line segments. This does not give a polygon as the the two extreme ends are still open. To complete the polygon, two additional imaginary class intervals, one before the smallest and another, after the largest class interval are taken with zero frequencies. The middle points of these additional class intervals are then joined to the extreme ends of the graph to give the required frequency polygon. Use of two additional class intervals on two extremes on the horizontal line helps in equalising the area under the frequency polygon and the histogram.

Frequency polygon can also be drawn without drawing histogram. Here the frequencies or the frequency densities (in case of unequal class intervals) are marked against the mid points of the class intervals and then all the adjacent points are joined consecutively by line segments. To complete the polygon, as before, two additional class intervals, one preceding the smallest class interval and the other succeeding the largest class interval are used with zero frequencies and then their middle points are located. To these points, the two extreme tails of the graph are then joined. This would give the required frequency polygon.

Advantages of Frequency Polygon

Frequency polygons are useful for graphical comparison of a number of frequency distributions because, on the same axis several frequency polygons can be drawn where as for each frequency distribution, a separate histogram is necessary. The frequency

Marks	No of Students
0-9	8
10-19	12
20-29	27
30-39	32
40-49	45
50-59	38
60-69	22
70-79	14
80-89	7
90-99	3
Total	208

Solution :

Fig - 11

HISTOGRAM AND FREQUENCY POLYGON OF MARKS

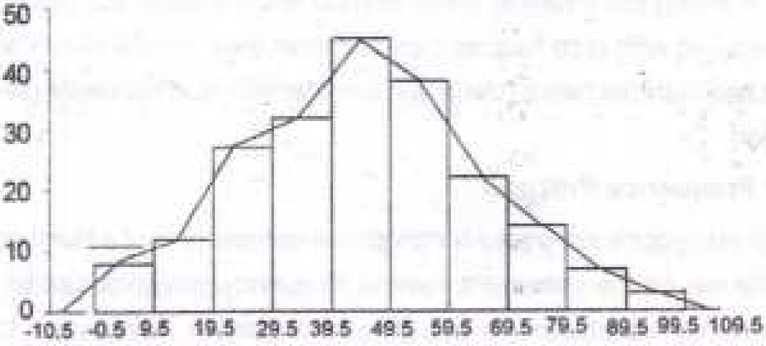
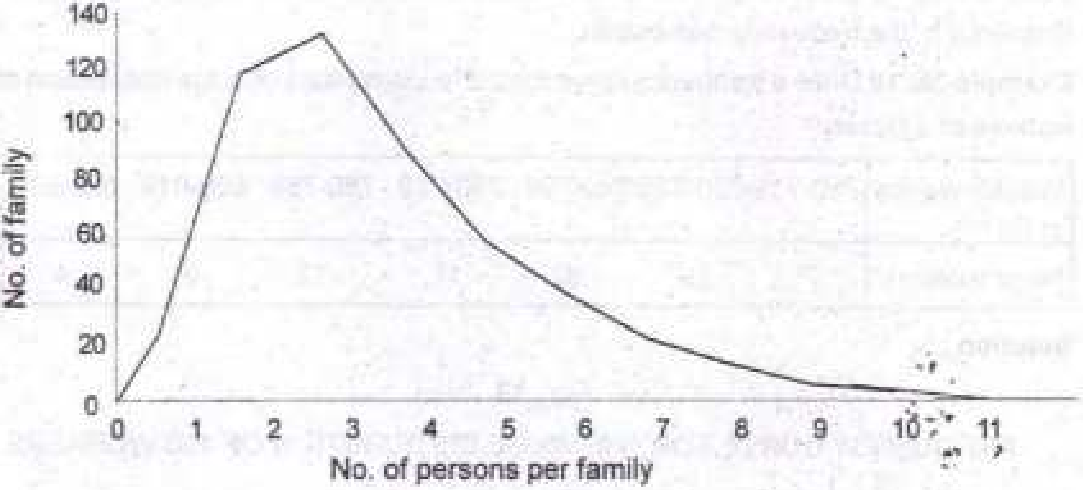


Fig - 12

FREQUENCY POLYGON OF NUMBER OF PERSONS PER FAMILY



3B.15 FREQUENCY CURVE

Very often frequency distributions are presented by curves, called frequency curves. These are smooth and free hand curves drawn through the vertices of frequency polygons. In a frequency curve the irregularities, if any, present in the frequency distribution are eliminated. The area under the frequency polygon (or histogram) is approximately equal to the area under the frequency curve. The difference between the frequency curve and the frequency polygon of a frequency distribution is that the frequency polygon contains angular points whereas the frequency curve is smooth and regular. But both of these have similar shape. A frequency curve may be regarded as a limiting form of the frequency polygon as the number of observations in the data become large and the class intervals are made small.

sequentially first by drawing a histogram, then a frequency polygon and finally the smooth curve. No doubt, a frequency polygon can be drawn without drawing a histogram. But in such a case, smoothing of the curve would be more subjective. The two extremes of the smooth curve end at the mid points of the two additional class intervals taken on either sides with zero frequency. The area under the frequency curve represents the total frequency of the frequency distribution.

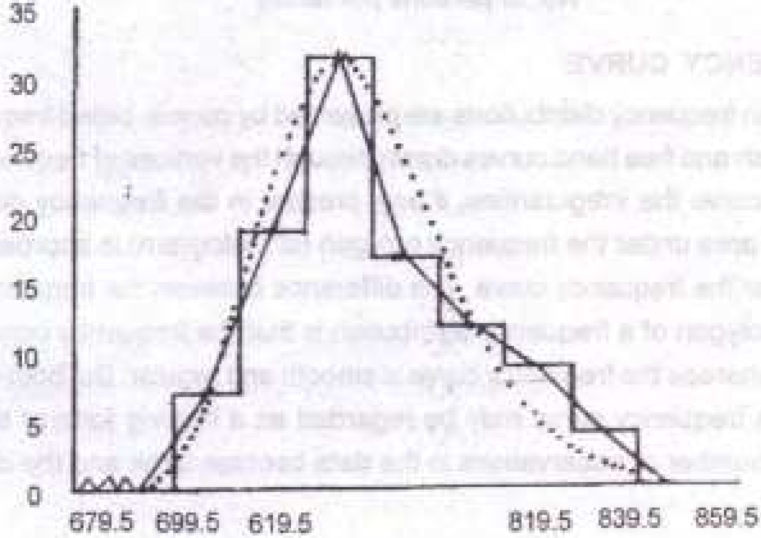
Example 3B.19 Draw a frequency curve for the following data of wage distribution of 100 workers of a factory.

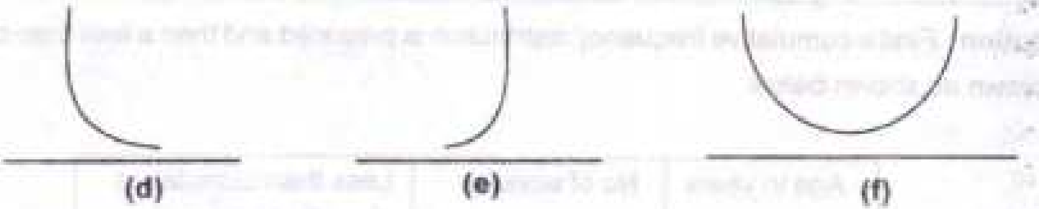
Weekly wages in Rs.	700-719	720-739	740-759	760-779	780-799	800-819	820-839	Total
No of workers	7	19	32	17	12	9	4	100

Solution :

Fig - 13

FREQUENCY CURVE FOR THE WAGE DISTRIBUTION OF 100 WORKERS.





Curves of the shape as in (a) are called symmetrical, curves of the shapes as in (b) and (c) are called moderately asymmetrical, curves of the shape as in (d) and (e) are called J-shaped and of the shape (f) U -shaped.

3B.16 CUMULATIVE FREQUENCY CURVE OR OGIVE

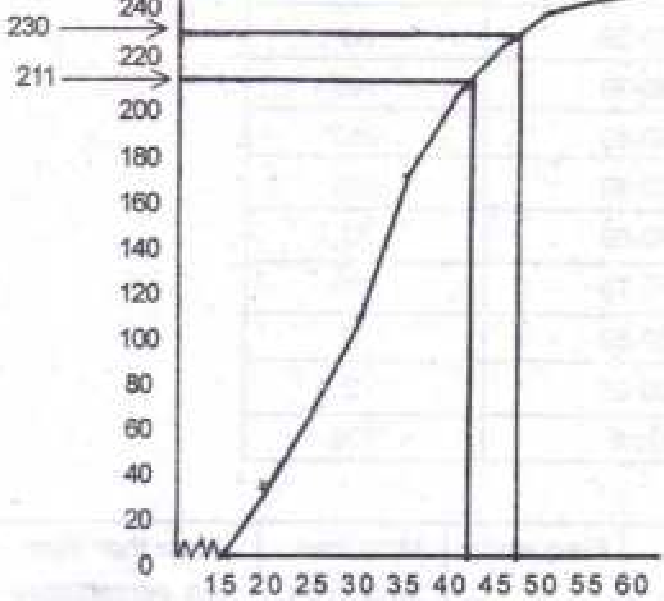
A cumulative frequency curve or an ogive (pronounced as ojiv) is graphical representation of the cumulative frequency distribution of a grouped data. It consists of plotting of cumulative frequencies against the class boundaries of the respective class intervals and then joining the plotted points by a curve. When the cumulative frequencies, from the lower end of the variable, are plotted against the upper boundaries of the respective class intervals, we have a less than ogive. Similarly when the frequencies are cumulated from the upper end and plotted against the lower boundaries, we get a more than ogive. A less than ogive is non-decreasing curve having an elongated S shape (   ) and a more than ogive is non-increasing having the shape of the mirror image of an elongated S (   ).

Both, the more than and less than ogives can be drawn on the same graph. They intersect at a point. They are useful, particularly, in locating the partition values of a frequency distribution graphically. They also can be used to compare two or more frequency distributions. If the class frequencies are large, they can be expressed as percentages of the total frequency and then these percentages of the cumulative frequencies can be used to obtain 'percentile curve'.

No of workers	28	35	42	64	33	22	16	6	4	250

Solution : First a cumulative frequency distribution is prepared and then a less than ogive is drawn as shown below.

Age in years	No of workers	Less than cumulative frequency
15-20	28	28
20-25	35	63
25-30	42	105
30-35	64	169
35-40	33	202
40-45	22	224
45-50	16	240
50-55	6	246
55-60	4	250
Total	250	



From the graph of the ogive in Fig-15 it is clear that

No. of workers whose age is less than 47 years = 230 and

No. of workers whose age is less than 42 years = 211

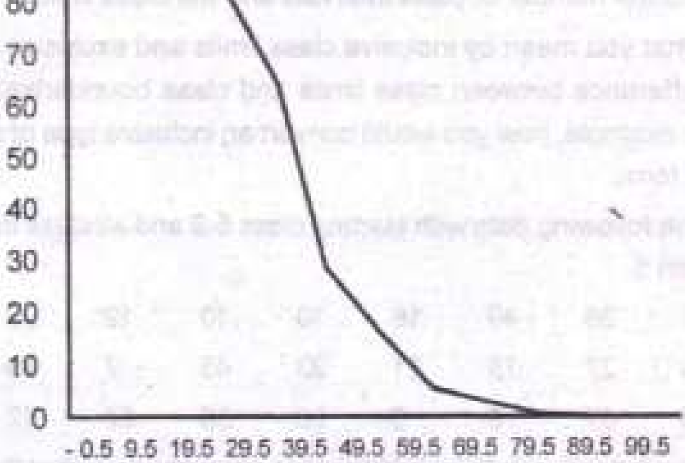
∴ The number of workers in the age group 42 – 47 is given by

$$230 - 211 = 19$$

20-29	63
30-39	145
40-49	240
50-59	103
60-69	72
70-79	21
80-89	5
90-99	2
Total	700

Solution :

Marks	Class boundaries	Frequency	More than cum. Freq	More than cum freq. percantages
0-9	−0.5 − 9.5	9	700	100.00
10-19	9.5 − 19.5	40	691	98.70
20-29	19.5 − 29.5	63	651	93.00
30-39	29.5 − 39.5	145	588	84.00
40-49	39.5 − 49.5	240	443	63.30
50-59	49.5 − 59.5	103	203	29.00
60-69	59.5 − 69.5	72	100	14.30
70-79	69.5 − 79.5	21	28	4.00
80-89	79.5 − 89.5	5	7	1.00
90-99	89.5 − 99.5	2	2	0.28
Total		700		



Exercises - 3B.1

1. What do you mean by classification of data? Discuss in brief, the various modes of data classification.
2. What are the objectives of classification? Write different types of classification.
3. Illustrate with suitable examples, the types of classification and write their uses.
4. What is meant by classification of data? State its important objectives. Briefly explain the different methods of classifying statistical data.
5. What are the purpose of classification of data? State the primary rules to be observed in classification. Write the advantages of classification.
6. State the principles underlying classification of data.
7. Distinguish between grouped and ungrouped frequency distributions. Explain the guide lines for the construction of a frequency distribution.
8. Discuss the problems in the construction of a frequeny distribution from raw data with particular reference to the choice of number of classes, the magnitude of the class intervals and the class limits.

referene to the number of class intervals and the class limits.

11. Explain what you mean by inclusive class limits and exclusive class limits. Bring out the difference between class limits and class boundaries. Explain, with the help of an example, how you would convert an inclusive type of classification to an exclusive form.

12. Classify the following data with starting class 5-9 and all class intervals having the same width 5.

12	36	40	16	10	10	19	20	28	30
19	27	15	21	33	45	7	19	20	26
26	37	6	5	20	30	37	17	11	20

13. Classify the following data by taking class intervals such that their mid values are 17, 22, 27, 32, etc.

30	42	30	54	40	48	15	17	51	42	25	41	33
30	27	42	36	28	26	37	54	44	31	36	40	41
36	22	30	31	19	48	16	42	32	21	22	46	21

14. The class marks of a grouped frequency distribution are 125, 132, 139, 146, 153, 160, 167, 174 and 181. Find (i) the magnitude of the class intervals (ii) the class boundaries and (iii) the class limits when the observations are rounded to the nearest integers.

15. Giving suitable examples, distinguish between :-

(i) Discrete and Continuous variable

(ii) Exclusive and Inclusive classification.

(iii) More than and Less than cumulative frequency tables.

(iv) Bivariate frequency tables.

16. Explain, with the help of an example, the cumulative frequency distribution, more than and less than cumulative frequencies.

than 12 and 35 get less than 15 marks.

19. Follwing data gives information about the number of flights from an airport in 60 days. Prepare a grouped frequency distribution and obtain the more than cumulative frequencies.

No of flights below :	5	10	15	20	25
No. of days :	18	22	35	41	60

20. Differentiate between continuous and discrete variables. Giving reasons, state whether the following variables are continuous or discrete.

(a) Age on the last birth day

(b) Body temperature of patients

(c) Length of rooms

(d) Number of share holders in a company

21. Prepare a frequency distribution of the words of the following extract according to their length (number of letters), omitting the punctuation marks. Also find (i) the number of words with 7 or less letters, (ii) the proportion of words with 5 or more letters, (iii) the percentage of words with not less than 4 or not more than 7 letters. "Success in the examination confers no absolute right to appointment, unless Government is satisfied, after such enquiry as may be considered necessary, that the candidate is suitable in all respects for appointment to the public service."

22. A company pays out bonus to its employees daily as under :

Daily salary (Rs) :	100-200	200-300	300-400	400-500	500-600	600-700
Daily bonus (Rs) :	10	20	30	40	50	60

Actual daily salaries of the employees, in rupees, are as under

175, 375, 225, 465, 375, 530, 478, 480, 525, 320

451, 345, 382, 471, 280, 450, 650, 515, 510, 225

Find the total daily bonus of the employees.

24. Following data give the marks of 24 students in two subjects Statistics (X) and Economics (Y). Taking class intervals 0-4, 5-9 etc. for both X and Y, consturct a bivariate frequency distribution.

(X,Y): (15,13), (2,2), (2,1), (3,7), (16,8), (4,9), (18,12), (5,9), (4,17), (17,16), (6,6), (19,18), (14,11), (9,3), (8,5), (13,4), (10,10), (13,11), (11,14), (11,7), (12,18), (18,15), (9,15), (17,3)

25. **Fill in the blanks.**

a) Variables are of two kinds .... and ....

b) Arrangement of data according to common characteristics is called ....

c) Data are classified on the basis of .... in chronological classification.

d) In .... classification the data are classified according to location.

e) Class mark (mid point) is the value lying halfway between ....

f) According to Sturges rule, the number of class intervals is given by k = ...

g) The magnitude of the class interval is the differene between ....

h) .... of data is a function similar to that of sorting letters in the post office according to destinations.

i) Class intervals can be .... and ....

j) In a grouped frequency distribution, the number of class intervals does not usually exceed ....

k) In exclusive class intervals, the class limits are equal to the class ....

l) Inclusive type of classes are preferred for .... variables.

m) Marks in a test and number of road accidents are two examples of ....

n) The classification where the lower limit of the smallest class or the upper limit of the largest class are not specified are called .... classes.

o) The difference between the upper limit and the lower limit of a class is called the .... of the class.

.... and ....

s) For a fequency distribution, the number of observations falling below the upper boundary of a particular class interval is called the .... frequency.

t) Sorting out on the basis of a single characteristic is called .... classification.

Excercises - 3B.2

1. Explain the terms classification and tabulation and pointout their importance in a statistical investigation. What precautions would ,ou take in tabulating statistical data?

2. Describe the chief functions of tabulation.

3. Explain the purpose of classification and tabulation of data. State the rules that serve as guide line in tabulation of data.

4. What do you mean by tabulation of data? What precautions would you take while tabulating data? Mention requisites of a good table.

5. Explain what you mean by tabulation. State the important points that should be kept in mind while tabulating data.

6. Write the rules to be followed in the preparation of a statistical table.

7. Briefly discuss the essential features of a statistical table.

8. Comment on the statement : "In collection and tabulation of data, commonsense is the chief requisite and experience is the chief teacher."

9. Explain and discuss the various types of tables used in statistical investigation.

10. The total number of accidents in Southern Railways in 1970 was 3500 and it decreased by 300 in 1971 and by 700 in 1972. The total number of accidents in metre-gauge section showed a progressive increase from 1970 to 1972. It was 245 in 1970, 346 in 1971 and 428 in 1972. In the metre-gauge section "Not

Enrolment in Commerce, Science, Law and Arts faculties in the session 1993-94 were 90,000; 60,500; 50,500 and 40,600 respectively. During the next two sessions the enrolment figures in the Science faculty were 65,210 and 70,300. while in Law and Arts faculties the corresponding figures were 55,000; 60,000 and 48,000; 50,000 respectively. In commerce faculty the enrolment during 1994-95 and 1995-96 were 1,15,000 and 1,22,250.

(a) Represent the information in tabular form.

(b) From your table calculate the increase in total enrolment from 1993-94 to 1995-96.

13. An investigation conducted by the education department in a public library revealed the following facts. Tabulate the information in a table.

"In 1990 the total number of readers was 46,000 and they borrowed some 16,000 books. In 2000, the number of books borrowed increased by 4000 and the borrowers by 5%.

The classification was on the basis of three sections : literature, fictions and illustrated news. There were 10,000 and 30,000 readers in the sections literature and fictions respectively in the year 1990. In the same year 2,000 and 10,000 books were lent in the sections illustrated news and fiction respectively. Marked changes were seen in 2000. There were 7,000 and 42,000 readers in the literature and fiction sections respectively. So also, 4,000 and 13,000 books were lent in the sections illustrated news and fictions respectively."

14. Draw a blank table to show five categories of workers: regular, seasonal, casual clerical and supervisory, classified as skilled and unskilled which are further sub-divided as male and female workers.

(e) In the collection and tabulation of data .... is the chief requisite and .... is the chief teacher.

(f) In a statistical table, the principal basis for the arrangement of captions and stubs in a systematic order are .... and ....

(g) Classification is the .... step in ....

(h) In a table, captions refer to .... and stubs refer to ....

(i) In a statistical table, the data are arranged in .... and ....

(j) In a statistical table .... should be avoided, especially in title and headings

(k) In a table necessary explanation of the columns/rows are given in the ....

Exercises - 3B. 3

1. What are the merits and demerits of diagrammatic representation of statistical data?

2. Describe the advantages of diagrams for presenting statistical data. Name the different types of diagrams commonly used and mention the appropriate situtations for use of such diagrams.

3. Describe the different types of diagrams which are used to represent statistical data to show the salient characteristics.

4. State the different methods used for diagrammatic representation of statistical data and indicate briefly the advantages and disadvantages of each one of them.

5. Point out the usefulness of diagrammatic representation of facts and explain the construction of some of the diagrams you know.

Food	200	250
Clothing	100	200
House rent	80	100
Fuel and light	30	40
Education	90	210
Total	500	800

8. Draw a pie-diagram for the following data :

Agriculture and Rural Development	12.9%
Irrigation	12.5%
Energy	27.2%
Industry and Minerals	15.4%
Transport, and Communication	15.9%
Social services	16.1%

9. Name the most appropriate diagram to be used for representing information in each of the following cases:

(a) Scientific manpower in four consecutive years in four different fields of science.

(b) Distribution of the number of children belonging to a school by their ages.

(c) The percentages of persons serving in different branches of the army such as infantry, artillary, signals etc.

(d) Production of wheat in India in each year from 1975 to 2004.

(e) The number of accidents on a given road between two locations during each of the last 20 days.

(f)' The number of male and female students in a college during eah of last six years.

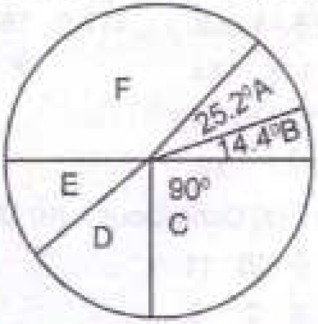
(g) The distribution of expenditure of a bank under different heads in a year.

1990 to 1997.

(m) Distribution of weights of 500 school children in the age group 15 – 18 years.

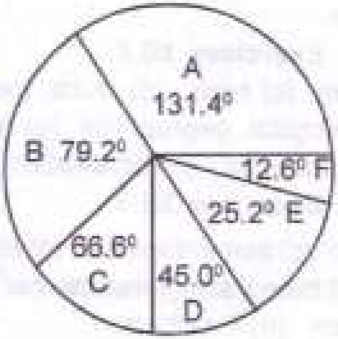
(n) Political partywise percentage of votes cast during a parliamentary election.

10. The following pie-diagram represents the relative shares of various sectors of development A, B, C, D, E and F in a country, under a five years plan.



Calculate the relative percentages of shares in A, B, C, D, E, F showing all the respective angles. If the total plan out lay is Rs. 16,000 crores, calculate the respective values of the shares in all the sectors individually. Present the calculated figures in a tabular form.

11. The following is a pie-diagram representing the values export of six commodities A, B, C, D, E and F in a country. If the total value of exports is Rs. 9,600 crores, calculate the values of exports of the commodities A, B, C, D, E and F respectively.



12.

C.I.:	5 – 9	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49
f :	3	4	6	5	4	3	3	1	1

13.

C.I.:	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49	50 – 54
f :	4	4	4	8	4	9	3	3

14. (i) 7 (ii) 121.5 - 128.5, 128.5 - 135.5, ..., 177.5 - 184.5

(ii) 122 – 128, 129 – 135, ..., 178 – 184

18.

C.I.:	0 – 3	3 – 6	6 – 9	9 – 12	12 – 15
f :	3	9	13	17	15

20. (a) Discrete (b) Continuous (c) Continuous (d) Discrete

21.

x :	2	3	4	5	6	7	8	9	10	11
f :	9	6	2	2	2	4	3	3	2	3

(i) 25 (ii) $\frac{16}{36} = 0.5278$ (iii) $\frac{10}{36} \times 100 = 27.28$

22. 720

25. (a) Discrete, Continuous, (b) Classification (c) Time (d) Geographical (e) Lower and Upper limit of the class (f) 1 + 322$\log_{10}$N (g) the upper and lower boundaries of the class (h) Classification (i) Inclusive and Exclusive (j) 15 (k) Boundaries (l) Discrete variable (m) Discrete variable (n) Open End (o) Width (p) 8 (q) Frequency (r) 44 – 52 (s) Less than Cumulative (t) Simple.

Exercises 3B.2

15. (a) classification (b) tabulation (c) body (d) stubs, captions (e) commonsense, experience (f) alphabetical, chronological, gegraphical (g) first, tabulation (h) column heading, row heading (i) rows and columns (j) abbreviations (k) footnote.

Exercises 3B.3

9. (a) subdivided bar (b) simple bar (c) percentage bar (d) simple line bar (e) simple line bar (f) multiple bar (g) Pie-chart (h) histogram (i) multiple bar (j) simple bar (k) simple line bar (l) multiple bar (m) hiostogram (n) Pie-chart.

* * *

co-operate.

(iii) The sphere of enquiry is very wide and the transportation cost is much.

(iv) The informants are literate and willing to co-operate.

(b) Which of the following is not a source of secondary data ?

(i) Census data used for purpose of National Income.

(ii) Data published by the Director of Economics and Statistics, Odisha and used by a research scholar.

(iii) Data collected by appointing local agents and used by an organisation.

(iv) Data collected by research scholars and submitted for award of degrees and used by a research organisation.

(c) Which of the following does not belong to one of the types of classification ?

(i) Geographical (Region wise) classification.

(ii) Chronological classification

(iii) Quantitative classification

(iv) Multiplicative classification

(d) Which of the following is correct ?

(i) Secondary data should be accepted at their face value without scrutiny.

(ii) There is no difference between primary data and secondary data.

(iii) Secondary data are always more dependable than primary data.

(iv) The difference between primary data and secondary data is one of degree only.

Classification is the process of arranging data in

(i) Different rows

(ii) Different columns

(iii) Different rows and columns

(iv) Grouping of related facts in different categories

(g) For a grouped frequency distribution, the number of classes should ideally be which of the following ?

(i) Less than 5 (ii) Greater than 8

(iii) Between 8 and 15 (iv) Greater than 20

(h) Which of the following is a one dimensional diagram ?

(i) Simple Bar diagram (ii) Diagrams represented by rectangles

(iii) Diagrams represented by circles (iv) Diagrams represented by cubes

(i) Which of the following diagrams is suitable to represent a frequency distribution ?

(i) Multiple Bar Diagram (ii) Histogram

(iii) Historigram (iv) Sub-divided Bar Diagram

(j) Which of the following is not true ?

(i) A frequency curve is a smooth free hand curve.

(ii) A frequency polygon and the corresponding frequency curve are the same.

(iii) The area under the frequency curve is approximately equal to the area under the frequency polygon.

(iv) The frequency curve and the frequency polygon for the same frequency distribution have similar shape.

(iv) To represent data relating to production of wheat in India for five different years, ________ diagram is suitable.

(v) Cumulative frequencies are required for obtaining ________ values.

(b) Write True (T) or False (F) for each of the following :

(i) Secondary data do not need much scrutiny and should be accepted at their face values.

(ii) Inclusive method of classification are used for obtaining partition values.

(iii) In a grouped frequency distribution the individual identity of the data is lost.

(iv) Exclusive method of classification is necessary to prepare tables.

(v) In chronological classification, data are obtained on the basis of locations.

3. Give short answers to the following questions :

(a) Indicate the difference between primary data and secondary data.

(b) State two limitations of secondary data.

(c) Name the various methods of collecting primary data.

(d) Write the difference between schedules and questionnaire methods of collecting primary data.

(e) Write the advantages of frequency polygon.

(f) Write the objectives of classification.

(g) Indicate the uses of more than and less than Ogives.

(h) Write the uses of diagrams in presenting statistical data.

(i) Write the purpose of tabulation of data.

(j) What is percentage Bar Diagram ?

(b) (i) F (ii) F (iii) T (iv) F (v) F

***

In the previous chapter we learned some techniques of presenting statistical data viz. classification, tabulation, diagrams, graphs and charts. Comparison of data through these techniques is based on subjective approach and hence can be interpreted by different persons in different ways as per their own judgements.

One of the methods of presenting data objectively is through concise summary figures, called measures. The most widely used measures are the measures of central tendency or measures of location or averages. An average is a single number which is considered to be the point of condensation of the total data and is located at a point around which most of the data values cluster. It is a value that lies between the two extreme values of the data. Since an average is a value that lies at the centre or towards the centre of the data, it is called measure of central tendency. It is also called measure of location.

Different persons have defined measures of central tendency in their own ways, all having the same meaning that it gives a bird's eye view to the huge data and serves as a representative figure of the data. We give below one definition due to Croxton and Cowden. They define

"An average value is a single value within the range of the data that is used to represent all of the values in the series. Since an average is some where within the range of the data, it is sometimes called a central value".

4.2. VARIOUS MEASURES OF CENTRAL TENDENCY

There are many different types of averages. The most commonly used averages are:

(v) Harmonic mean

4.3. REQUISITES OF AN IDEAL MEASURE OF CENTRAL TENDENCY (AVERAGE)

Prof. Yule states that the following characteristics are to be satisfied by an ideal measure of central tendency.

(a) It should be rigidly defined.

(b) It should be based on all the observations of the data.

(c) It should be simple to understand and easy to calculate.

(d) It should be capable of further algebraic treatments. i.e. the mathematical formula should be mathematically amenable to numerical manipulations.

(e) It should not be affected much by extreme items, i.e. should not give unduly great importance to the larger or the smaller items of the data.

(f) It should be least affected by fluctuations of sampling.

4.4. ARITHMETIC MEAN

Arithmetic mean of a set of observations is given by the sum of the observations divided by the total number of the observations. Thus, the arithmetic mean of marks of five students (50,52,57,52,53) = $\frac{1}{5}$ (50+52+57+52+53) = 52.8 marks.

In general, let

$X_1, X_2, \ldots\ldots X_n$

denote the numerical values of 'n' observations. The arithmetic mean of these n observations, denoted by $\overline{X}$, is given by

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (4.1)$$

where, $X_1, X_2, ..., X_n$ are the different values of the variable and $f_1, f_2, ..., f_n$ are the corresponding frequencies, the arithmetic mean is given by :

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i \qquad (4.2)$$

where, $N = \sum_{i=1}^{n} fi$

For a grouped frequency distribution, the class marks (or the mid points) of the class intervals are taken as the values of X_i.

Example 4.1. Calculate the arithmetic mean of the wages of 100 workers of a firm from the following frequency distribution.

Wages in Rs	80	82	85	90	95	98	100	110	Total
No. of Workers	5	8	17	30	18	11	8	3	100

Solution:

x	f	fx
80	5	400
82	8	656
85	17	1445
90	30	2700
95	18	1710
98	11	1078
100	8	800
110	3	330
Total	100	9119

Example 4.2. The marks of 200 school students is given in the following frequency distribution. Calculate the average mark.

Marks	11-20	21-30	31-40	41-50	51-60	61-70	71-80	Total
No. of Students	8	12	35	63	45	27	10	200

Solution :

Marks	**Class marks** **X**	**No. of Students** **f**	**fX**
11-20	15.5	8	124.0
21-30	25.5	12	306.0
31-40	35.5	35	1242.5
41-50	45.5	63	2866.5
51-60	55.5	45	2497.5
61-70	65.5	27	1768.5
71-80	75.5	10	755.0
Total		200	9560.0

Average, i.e. the arithmetic mean, $\overline{X} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$

$$= \frac{9560}{200} = 47.8 \text{ marks}$$

some authors.

Let the original frequency distribution be given by (x_i, f_i), $i = 1,2,3,\ldots,n$. So the arithmetic mean, by (4.2), is given by

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i, \text{ where , } N = \sum_{i=1}^{n} f_i$$

$$\text{Let } d_i = \frac{x_i - A}{h} \tag{4.3}$$

where, A is any arbitrary value and $h \neq 0$.

$$\therefore \quad x_i = A + h\,d_i \tag{4.3 a}$$

Multiplying by fi and summing both sides of (4.3a) over $i = 1,2,\ldots,n$ we have,

$$\sum_{i=1}^{n} f_i x_i = \sum_{i=1}^{n} f_i (A + h\,d_i)$$

Dividing by N we get

$$\frac{1}{N}\sum_{i=1}^{n} f_i x_i = \frac{1}{N}\sum_{i=1}^{n} A f_i + \frac{1}{N}\sum_{i=1}^{n} h\,f_i\,d_i$$

$$\text{i.e} \quad \overline{X} = A.\frac{1}{N}\sum_{i=1}^{n} f_i + h\frac{1}{N}\sum_{i=1}^{n} f_i\,d_i$$

$$= A.\frac{1}{N}.N + h.\overline{d}$$

$$\text{i.e} \quad \overline{X} = A + h\overline{d} \tag{4.4}$$

where $\overline{d} = \frac{1}{N}\sum_{i=1}^{n} f_i\,d_i$ is the arithmetic mean of d_i values.

Note 2. Although, theoretically , there is no restriction in the choice of the values of 'A' and 'h', as a thumb rule , 'A' should be taken a value at or around the middle of the data values and 'h' should be taken equal to the common magnitude of the class intervals. In case, the class intervals are not of equal magnitude, the HCF of the class intervals may be taken as the value of 'h' or 'h' may be taken as1.

Example 4.3. Find tl e average of the salaries of 210 persons of a firm given in the following frequency distribution.

Monthly salary in Rs.	No of persons
3000-3999	15
4000-4999	13
5000-5999	46
6000-6999	67
7000-7999	32
8000-8999	23
9000-9999	14
Total	**210**

4000-4999	13	4499.5	−2	−26	−117
5000-5999	46	5499.5	−1	−46	
6000-6999	67	6499.5	0	0	
7000-7999	32	7499.5	1	32	120
8000-8999	23	8499.5	2	46	
9000-9999	14	9499.5	3	42	
Total	**210**			**3**	

The average of the salary or the arithmetic mean is given by

$$\overline{X} = A + h.\overline{d}$$

$$= A + h.\frac{1}{N}\sum_{i=1}^{n} f_i\, d_i$$

$$= 6499.5 + 1000x\frac{3}{210}$$

= 6499.5+14.28

= 6513.78 Rs. approximately.

4.4.2. Algebraic properties of Arithmetic mean

P. 1. The sum of deviations of the observations of a data from its arithmetic mean is zero.

Let (x_i, f_i), i = 1, 2,....... n represent a frequency distribution and let $\overline{X}$ be its arithmetic mean. This property states that,

$$= \sum_{i=1} f_i x_i - \bar{X} \sum_{i=1} f_i$$

$$= N\bar{X} - N\bar{X}$$

$$= 0 \qquad\qquad (\because \ \Sigma f_i = N)$$

P.2. Let $\bar{X}$ be the arithmetic mean of a set of data having N_1 observations and $\bar{Y}$ be the arithmetic mean of another set of data having N_2 observations. The arithmetic mean M, of the combined data of (N_1+N_2) observations is given by

$$M = \frac{N_1\bar{X} + N_2\bar{Y}}{N_1 + N_2} \tag{4.5}$$

In general, for k different data sets having arithmetic means $\bar{X}_i$ and corresponding number of observations N_i (i = 1,2,3,....k) the combined mean M, is given by

$$M = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + ... + N_k\bar{X}_K}{N_1 + N_2 + ... + N_K} \tag{4.5a}$$

We prove the property for two sets of data, which can be extended to k distributions like wise.

Proof: Let (x_i, f_i) i = 1,2,3,...,m, $\sum_{i=1}^{m} f_i = N_1$ be the frequency distribution of one series of observations and (y_j, l_j), j = 1,2,...,n, $\sum_{j=1}^{n} l_j = N_2$ be the frequency distribution of another series of observations.

$$\sum_{i=1}^{m} f_i x_i + \sum_{j=1}^{n} l_j y_j$$

and the total number of observations would be $(N_1 + N_2)$.

Thus, the arithmetic mean M, of the combined distribution would be

$$M = \frac{\sum_{i=1}^{m} f_i x_i + \sum_{j=1}^{n} l_j y_j}{N_1 + N_2} = \frac{N_1 \overline{X} + N_2 \overline{Y}}{N_1 + N_2}$$

The arithmetic mean of the combination of k component series can be found by using a similar technique.

Example. 4.4. In a class with two sections A and B, the average of the marks in Statistics of the students of section A is 62 and that of section B is 60. If the number of students in these sections be 80 and 70 respectively, find the average of the marks of all the 150 students taken together.

Solution :

Let $\overline{X}$ and $\overline{Y}$ be the average of the marks of the students of sections A and B with respective number of students N_1 and N_2. So, $\overline{X} = 62$, $\overline{Y} = 60$, $N_1 = 80$ and $N_2 = 70$

Average marks of all 150 students taken together is given by,

$$M = \frac{N_1 \overline{X} + N_2 \overline{Y}}{N_1 + N_2} = \frac{80x62 + 70 \times 60}{80 + 70}$$

$$= \frac{4960 + 4200}{150} = \frac{9160}{150} = 61.067 \text{ marks appr.}$$

We know that $M = \frac{N_1X + N_2Y}{N_1 + N_2}$

Given : $M = 600$, $\overline{X}$ =630 and $\overline{Y} = 530$

$$\therefore\ 600 = \frac{N_1 \times 630 + N_2 \times 530}{N_1 + N_2}$$

i.e. $600\ N_1 + 600\ N_2 = 630\ N_1 + 530\ N_2$

i.e. $30\ N_1 = 70\ N_2$

i.e. $\frac{N_1}{N_2} = \frac{70}{30} = \frac{7}{3}$

So, the percentage of male workers $= \frac{7}{7+3} \times 100 = 70$

and the percentage of female workers $= \frac{3}{7+3} \times 100 = 30$

Hence, there are 70% male workers and 30% female workers in the factory.

Example.4.6. The average height of 149 students out of 150 is 162.9 cms. The height of the 150th student is 14.9 cms more than the average height of all the 150 students taken together. Find the height of the 150 th student and the average height of all the150 students.

Solution: Let the average height of all the 150 students be denoted by $\overline{X}$.

As the average height of 149 students is 162.9, the total height of 149 students is = 149 x 162.9 = 24272.1 cms. and the height of the 150th student is $\overline{X}$ + 14.9.

Now, $\overline{X} = \frac{\text{Total height of 150 students}}{150}$

$$\therefore \quad 150\,\overline{X} = 24287 + \overline{X}$$

$$\text{i.e. } 149\,\overline{X} = 24287$$

$$\text{i.e. } \overline{X} = \frac{24287}{149} = 163 \text{ cms}$$

So, the height of the 150th student is 163 + 14.9 = 177.9 cms.

P.3. The sum of squares of differences of the observations of a series taken from its arithmetic mean is less than the sum of squares of differences taken from any other value.

Let $\overline{X}$ be the arithmetic mean of n observations $X_1, \ldots, X_n$ and let A be any arbitrary value. We have to show that

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 \leq \sum_{i=1}^{n}(X_i - A)^2 \qquad (4.6)$$

$$\text{LHS} \quad = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - A + A - \overline{X})^2$$

$$= \sum_{i=1}^{n}\{(X_i - A) - (\overline{X} - A)\}^2$$

$$= \sum_{i=1}^{n}\{(X_i - A)^2 + (\overline{X} - A)^2 - 2(X_i - A)(\overline{X} - A)\}$$

$$= \sum_{i=1}^{n}(X_i - A)^2 + n(\overline{X} - A)^2 - 2(\overline{X} - A)\sum_{i=1}^{n}(X_i - A)$$

$$= \sum_{i=1}^{n}(X_i - A)^2 + n(\overline{X} - A)^2 - 2(\overline{X} - A).n(\overline{X} - A)$$

$$= \sum_{i=1}^{n}(X_i - A)^2 + n(\overline{X} - A)^2 - 2n(\overline{X} - A)^2$$

$$= \sum_{i=1}^{n}(X_i - A)^2 - n(\overline{X} - A)^2$$

$(X_i, f_i), i = 1, 2, \ldots, n.$

$$\text{i.e.} \sum_{i=1}^{n} f_i(X_i - \overline{X})^2 \le \sum_{i=1}^{n} f_i(X_i - A)^2 \qquad (4.6a)$$

Note.2. Relation (4.6) can be stated otherwise as 'the sum of squares of deviations taken from the arithmetic mean is least'.

4.4.3. Merits and Demerits of Arithmetic mean.

Merits:

1. It is rigidly defined.
2. It is based on all the observations of the data.
3. It is amenable to further algebraic treatments. The mean of a combined series of distributions can be found in terms of the means and the number of observations of the distributions. Further, arithmetic mean found by using one or more incorrect observations can be replaced by correct observations and the correct value of the arithmetic mean can be calculated from incorrect value without knowing the values of all the observations in the data.
4. In comparison with other averages, it is least affected by fluctuations of sampling.
5. It is not difficult to calculate and has an intuitive appeal as a typical value.

Demerits:

1. It is not that easy to calculate like some of the other averages. It can neither be determined graphically nor by a mere inspection.
2. Since it is based on all the observations of the data, in the absence of even a single value of the observations, its calculation is not possible. Such situations occur frequently for primary data where an observation is either incorrect or missing or illegible. One can avoid such situations either by omitting the observation

observations. For example, if we want to compare the standards of two students A and B on the basis of their average performance in three subjects, it may so happen that their average scores are equal but their individual scores differ widely. One might fail in one subject but might have a higher average score than the other student.

6. It may lead to results which are not easily understood. For example, we may find that the average number of children per family in a locality is 1.7; which does not correspond to a real observation.

7. It gives equal importance to all the observations of the data. For example, the arithmetic mean of five observations 8,9,10,11,12 is 10. If we replace 10 by a large value like 30, the arithmetic mean becomes 14 which is larger than four of the five observations.

Example 4.7. Show that, if all the observations of a data are added, subtracted, multiplied or divided by a non zero constant, the arithmetic mean is also added, subtracted, multiplied or divided by the same constant.

Solution : Let $X_1, X_2, \ldots\ldots X_n$ be the 'n' observations of a data and let $\overline{X}$ be their arithmetic mean i.e.

$$\therefore \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Assume that $c \neq 0$, is a constant.

Now, let

$$U_i = X_i + c,\ V_i = X_i - c,\ W_i = cX_i \text{ and } Z_i = \frac{X_i}{c}$$

$$\text{So, } \overline{U} = \frac{1}{n}\sum_{i=1}^{n} U_i \quad = \frac{1}{n}\sum_{i=1}^{n}(X_i + c) \quad = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{nc}{n} \quad = \overline{X} + c$$

$n\ i{=}1 \qquad n\ i{=}1\ c \qquad c\ n\ i{=}1 \qquad c$

Thus we find that, if

$U = X + c, \qquad \overline{U} = \overline{X} + c$

$V = X - c, \qquad \overline{V} = \overline{X} - c$

$W = cX, \qquad \overline{W} = c\overline{X}$

and $\quad Z = \frac{X}{c}, \qquad \overline{Z} = \frac{\overline{X}}{c}$

Example 4.8. A student calculated the arithmetic mean of 100 observations as 82. At the time of checking, it was found that he had wrongly copied two observations 29 and 62 as 92 and 26. Find the correct value of the arithmetic mean.

Solution : No. of observations, N = 100

Incorrect mean, $\overline{X}' = 82$

Incorrect observations are 92 and 26.

Correct observations are 29 and 62.

In correct total $\Sigma X' = 100 \times 82 = 8200$

Correct total $\Sigma X = 8200 - (92+26) + (29+62)$

$= 8200 - 118 + 91$

$= 8173.$

Correct Arithmetic Mean, $\overline{X} = \frac{8173}{100} = 81.73.$

Example 4.9. The following data represents an incomplete grouped frequency distribution of inclusive class intervals where X denotes the mid point of the class intervals and C, a nonzero constant.

X – C = d	f	fd	
– 15	3	– 45	–375
– 10	18	– 180	
– 5	30	– 150	
0	45	0	
5	28	140	435
10	19	190	
15	7	105	
Total	**150**	**60**	

Since $X - C = d$, $\bar{d} = \bar{X} - C$

i.e. $C = \bar{X} - \bar{d}$

Now, $\bar{d} = \frac{1}{N}\sum_{i=1}^{n} f_i d_i = \frac{1}{150} \times 60 = 0.4$

$\therefore \quad C = \bar{X} - \bar{d} = 52.4 - 0.4 = 52$

From the given data it is clear that the magnitude of the class intervals are all equal to 5. The class mark of the class interval having value of X – C = 0 would be 52. All other class marks and the class limits can now be written as in the following table.

52	50-54
57	55-59
62	60-64
67	65-69

50-54	45
55-59	28
60-64	19
65-69	7
Total	150

4.4.4. Weighted Arithmetic Mean:

In the calculation of arithmetic mean of either a frequency distribution or of ungrouped data, each observation has been given equal importance. Such an average is called simple 'arithmetic mean'. But on many occasions, all the observations of the data are not of equal importance. In such cases, some of the observations are given more importance than the others. Here, instead of simple arithmetic mean, a weighted arithmetic mean is computed by assigning due importance to the specific values of the data. The weighted means thus computed become more representative than the unweighted ones.

Illustration: Suppose a student's average score is to be determined from the percentage of marks scored by him in two internal examinations and one final examination. It is decided to give 20% importance to each of the scores of internal examinations and 60% importance to the scores of the final examination. To determine his average score, weighted mean is to be computed.

Let $X_1, X_2, \ldots X_n$ be n observations of a data with corresponding weights $W_1, W_2, \ldots W_n$. We define the weighted arithmetic mean of the data, $\overline{X}_w$ as

$$\overline{X}_w = \frac{\sum_{i=1} W_i f_i X_i}{\sum W_i f_i} \qquad (4.7a)$$

Note: The formula (4.2), in a sense, may be treated as the weighted mean with the frequencies as the weights i.e. all observations have equal weights.

If the weights are equal for all the observations of the data, the simple arithmetic mean would be equal to the weighted arithmetic mean. If greater weights are given to larger observations and smaller weights to smaller observations then the weighted mean would be greater than the simple mean. Similarly, if smaller weights are given to larger observations and greater weights to smaller observations, then the weighted mean would be less than the simple mean.

Example 4.10. For entry into an Engineering course in an institution, only the marks of the students in the CHSE examination were considered. The average marks of the students were computed by giving a weight of 60 to each of the subjects Physics, Chemistry and Mathematics and weight 40 to each of the subjects English and MIL. A student's percentage of marks are given below. Find the weighted mean and the simple mean of his marks.

Subject	Percentage of marks
English	62
MIL	80
Physics	92
Chemistry	73
Mathematics	95
Total	**402**

Physics	92	60	5520
Chemistry	73	60	4380
Mathematics	95	60	5700
Total	**402**	**260**	**21,280**

The weighted mean, $\overline{X}_w = \dfrac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i} = \dfrac{21280}{260}$

$= 81.846\ \%$

Simple mean, $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i = \dfrac{402}{5} = 80.4\%$

Example 4.11. Find the simple arithmetic mean and the weighted mean of the first 'n' natural numbers where the weights are equal to the corresponding numbers.

Solution:

X	W	WX
1	1	1^2
2	2	2^2
3	3	3^2
4	4	4^2
.	.	.
.	.	.
n	n	n^2

$$\bar{X}_w = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i} = \frac{1^2 + 2^2 + ... + n^2}{1 + 2 + ... n} = \frac{2n(n+1)(2n+1)}{6n(n+1)} = \frac{2n+1}{3}$$

Note that for n >1, the weighted mean is greater than the simple mean. This is because, greater weights have been assigned to larger values.

4.5. MEDIAN

Median is one of the measures of central tendency. Median of statistical data set is the value of the middle item when the observations of the data are arranged either in ascending or in descending order of their magnitudes. It is that value of the series below which there lie exactly 50% of the observations. For a grouped frequency distribution, the median divides the group in to two equal halves. 50% of the area under the frequency curve lie below the ordinate at the median and other 50% of the area lie above the ordinate. In the words of L.R. Corner. "The median is that value of the variable which divides the group into two equal parts, one part comprising all the values greater and the other, all the values less than the median".

4.5.1. Median for Ungrouped Data

For ungrouped data, the observed values of the variable are first arranged in ascending order (or in descending order)of their magnitudes. Then the middle observation is found which is the value of the median. The middle observation is the $\frac{N+1}{2}$ nd observation of the data, where N is the total number of observations.

observations at $\frac{N}{2}$nd and $\left(\frac{N}{2}+1\right)$st positions is taken as the value of the median.

Example 4.12. Find the median of the following marks of 10 students in a class.

Marks: 12,15, 8,12,10,16,9,17,14,11

Solution : Arranging the marks in ascending order of magnitude, we have

8,9,10,11,12,12,14,15,16,17.

Since, N = 10, median refers to the average of $\frac{10}{2}$ = 5 th and $\frac{10}{2}+1$ = 6 th observations i.e.

$$\text{Median} = \frac{12+12}{2} = 12 \text{ marks.}$$

4.5.2. Median for Frequency Distribution

In case of frequency distribution, the observations are already arranged in order of their magnitudes. Location of the median value is done by calculating the cumulative frequencies as is given in the following example.

Example. 4.13. The following data gives the number of days on which (X) students were absent from a class during a certain period. Calculate the median number of absentees.

No. of Students remaining absent : (X)	1	2	3	4	5	6	7	8	Total
No of days	5	15	22	11	2	6	3	1	65

3	22	42
4	11	53
5	2	55
6	6	61
7	3	64
8	1	65
Total	**65**	

Median = $\frac{N+1}{2}$ nd observation

$= \frac{65+1}{2}$ =33rd observation

Looking at the cumulative frequency column, we observe that the 33rd observation is 3 because 3 occupies all the days from 21st to 42 nd.

$\therefore$ The median number of the absentees = 3

4.5.3. Median for Grouped Frequency Distribution.

In a grouped frequency distribution of a continuous variable, the value of median is obtained by using linear interpolation. Here it is assumed that ordinate at the median divides the entire area under the histogram in to two equal halves. So the value of the variable corresponding to the number $\frac{N}{2}$ (not $\frac{N+1}{2}$ as in case of ungrouped data) in the cumulative frequency column would be the value of the median. The class, cumulative frequency of which just exceeds N/2, is called the median class and it contains the median.

F is the cumulative frequency of the class preceding the median class and

N is the total frequency.

Formula (4.8) is based on uniform distribution of the frequency over the median class.

Example 4.14. Calculate the median height for the following frequency distribution.

Heights in cms	No. of persons
140-144	9
145-149	21
150-154	33
155-159	48
160-164	62
165-169	41
170-174	24
175-179	12
Total	250

149.5-154.5	33	63
154.5-159.5	48	111
159.5-164.5	62	173
164.5-169.5	41	214
169.5-174.5	24	238
174.5-179.5	12	250
Total	**250**	

Now, $N = 250$

So, $\frac{N}{2} = \frac{250}{2} = 125$

i.e. the median refers to the value of the variable corresponding to the 125th position. Looking at the cumulative frequency column we find that cumulative frequency 125 is included in the class interval 159.5 – 164.5

$\therefore \quad l_1 = 159.5, \quad l_2 = 164.5, \quad f = 62$ and $\quad F = 111$

Substituting these values in (4.8) we get,

Median = 159.5 + (164.5 – 159.5)(125 – 111) /62

= 159.5 + (5 x 14)/62 = 159.5 + 1.129 = 160.629(cms)

4.5.4. Derivation of the Interpolation Formula for Median.

Consider the part of the cumulative frequency polygon containing the median of the frequency distribution of a continuous variable X. Let l_1 and l_2 denote respectively the lower and upper boundaries and f, the frequency of the class containing the median (or

We make the following assumption:

All the values of the variable in the median class are uniformly spread over the entire class interval.

From the Fig (17), it is clear that

$$\text{Median} = l_1 + AK$$

In the similar triangles DGH and DCE

$$\frac{GH}{CE} = \frac{DH}{DE}$$

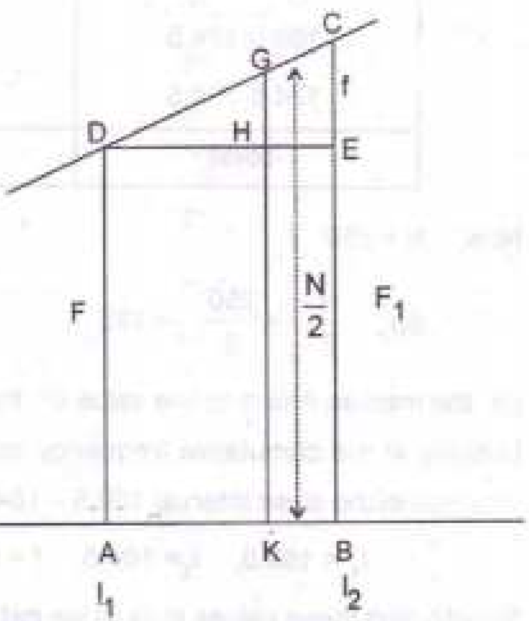
$$\text{i.e.}\quad \frac{KG - KH}{BC - BE} = \frac{AK}{AB}$$

$$\text{i.e.}\quad \frac{\frac{N}{2} - F}{f} = \frac{AK}{l_2 - l_1}$$

$$\text{So,}\quad AK = (l_2 - l_1)\left(\frac{\frac{N}{2} - F}{f}\right)$$

$$\text{Hence, Median} = l_1 + (l_2 - l_1)\frac{\frac{N}{2} - F}{f}$$

Fig – 17



Example 4.15. Calculate the median for the following frequency distribution representing the weights of 304 apples.

180-189	64
190-199	48
200-209	31
210-219	26
220-229	12
230 and above	8
Total	304

Solution:

Weights in gms	Class boundaries	No. of apples	Cum. Freq.
Less than 150	Less than 149.5	25	25
150-159	149.5-159.5	18	43
160-169	159.5-169.5	30	73
170-179	169.5-179.5	42	115
180-189	179.5-189.5	64	179
190-199	189.5-199.5	48	227
200-209	199.5-209.5	31	258
210-219	209.5-219.5	26	284
220-229	219.5-229.5	12	296
230 and above	229.5 and above	8	304
Total		304	

$\therefore \quad l_1 = 179.5, \quad l_2 = 189.5, \quad f = 64, \quad F = 115$

Using (4.8) we get,

$$\text{Median} = 179.5 + (189.5 - 179.5)\left(\frac{152-115}{64}\right)$$

$$= 179.5 + 10\text{x}\frac{37}{64}$$

$$= 179.5 + 10 \times 0.578$$

$$= 179.5 + 5.78$$

$$= 185.28 \text{ gms}$$

4.5.5. Merits and Demerits of Median.

Merits.

(i) It is rigidly defined except for 2n discrete values. In this case, the median can be defined by convention.

(ii) It is simple to understand and easy to calculate. It can also be located graphically and in some cases, by mere inspection.

(iii) It can be calculated for distributions having open end class intervals.

(iv) It is not affected by observations having extreme values.

(v) It is useful to find the average of data which cannot be precisely measured.

Demerits:

(i) It is not based on all the observations of the data. As such, the medians of data differing from one another in a number of ways can be the same. For example, the

(ii) It cannot be subjected to further algebraic treatments.

(iii) It is affected much by fluctuations of sampling as compared with the arithmetic mean, for small samples in particular.

(iv) In case of ungrouped distribution when the total number of observations is even, it does not refer to a particular observation of the data in which case, the arithmetic mean of the two middle observations is taken as the value of the median.

Uses:

(i) It is used to find the average of ordinal data which cannot be measured quantitatively but still can be arranged in a sequence. It can be used in distributions having open end class intervals.

(ii) It is used as an average for typical values like wages, distribution of wealth etc. where extreme values occur.

4.6. MODE

Mode is the most typical value of a data i.e. a value which occurs largest number of times and around which observations of the data cluster densely. In the words of Croxton and Cowden. "The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded as the most typical of series of values". In other words, mode is the value having greatest frequency density in its neighbourhood.

Consider the following statements:

(i) A student staying in a college hostel spends, on an average, Rs. 1500.00 pm.

(ii) The average size of shirts sold in a garment shop is 42.

the largest number of shirts sold have the size 42; while in the third statement we mean that the largest number of commuters in Bhubaneswar use motor bikes.

4.6.1. Computation of mode in Ungrouped Data.

Mode can sometimes be obtained by mere inspection of the frequencies of a frequency distribution . It is a value of the data that corresponds to the largest frequency. Consider the following frequency distribution

X	1200	1300	1400	1500	1600	1700	1800	Total
f	2	8	12	23	18	11	5	79

Here the maximum frequency is 23 which corresponds to the value 1500. So the mode is 1500.

This method of locating mode is applicable only if the frequency distribution is regular i.e. the frequencies gradually increase, reach a maximum and then decrease. But if any one or more of the following situations arise, the mode is determined by the method of grouping.

(i) The maximum frequency occurs for two or more values of the frequency distribution.

(ii) The maximum frequency occurs near the beginning or towards the end of the frequency distribution.

(iii) The frequency distribution is not regular.

4.6.2. Method of Grouping.

The frequencies of the values of the variables are added in two's and in three's starting from the beginning of the frequency distribution. Grouping in two's can be done in

Solution: We notice that the maximum frequency here is 40 and it occurs towards the beginning of the distribution. So, mode cannot be determined by mere inspection. To find the mode we follow the grouping method.

Grouping Table

X	f	Groups in two's		Groups in three's		
	(1)	(2)	(3)	(4)	(5)	(6)
26	12	33		73		
27	21		61		88	
28	40	67				102
29	27		62	90		
30	35	63			79	
31	28		44			55
32	16	27		31		
33	11		15			
34	4					
Total	**194**					

28	1	1			1	1	4
29		1	1	1	1	1	5
30			1	1		1	3
31				1			1
32							
33							
34							

From the above analysis table, it appears that the size 29 occurs largest number of times. So the size 29 is the modal size.

4.6.3. Mode for Grouped Data

For a grouped frequency distribution, mode is determined by using the following interpolation formula.

$$\text{Mode} = l_1 + (l_2 - l_1)\frac{f_m - f_0}{(f_m - f_0) + (f_m - f_1)}$$

$$= l_1 + (l_2 - l_1)\frac{f_m - f_0}{2f_m - f_0 - f_1} \qquad (4.9)$$

where, l_1, l_2, f_m are respectively the lower boundary, upper boundary and the frequency of the modal class (the class interval containing the mode), f_0 and f_1 are respectively the frequencies of the classes preceding and following the modal class.

In case, the maximum frequency does not correspond to the modal class, $(2f_m - f_0 - f_1)$ may be negative or zero. Such situations can be avoided by using the formula.

$$\text{Mode} = l_1 + (l_2 - l_1)\frac{|f_m - f_0|}{|f_m - f_0| + |f_m - f_1|} \qquad (4.9a)$$

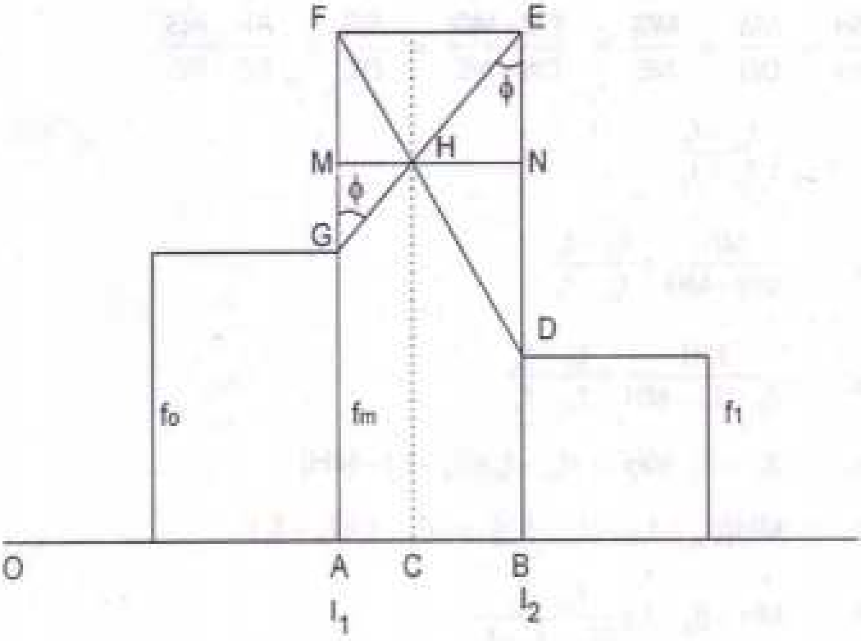
magnitudes. In case of unequal class intervals, these three consecutive class intervals are to be made equal under the assumption that all the observations are uniformly distributed over the corresponding class interval. Otherwise, the value of mode obtained by using (4.9) may give misleading results.

(iii) When the maximum frequency occurs for more than one class intervals or the maximum frequency occurs near the beginning or towards the end of the frequency distribution or the frequency distribution is not regular, grouping of the frequencies is necessary to locate the modal class. In such a situation (4.9a)is to be used.

4.6.4. Derivation of Interpolation Formula for Mode

Consider a frequency distribution having exclusive class intervals of a continuous variable and the portion of its histogram containing the mode.

Fig – 4.18



In the Fig 4.18,

$$\frac{FM}{MH}=\frac{DN}{NH}$$

i.e. $$\frac{MH}{NH}=\frac{FM}{DN} \qquad (4.9c)$$

Again, $$\frac{MH}{MG}=\frac{NH}{NE}$$

i.e. $$\frac{MH}{NH}=\frac{MG}{NE} \qquad (4.9d)$$

Thus, from (4.9c) and (4.9d),

$$\frac{MH}{NH}=\frac{FM}{DN}=\frac{MG}{NE}=\frac{FM+MG}{DN+NE}=\frac{FG}{DE}=\frac{AF-AG}{BE-BD}$$

$$=\frac{f_m-f_o}{f_m-f_1}$$

i.e. $$\frac{MH}{MN-MH}=\frac{f_m-f_o}{f_m-f_1}$$

i.e. $$\frac{MH}{(l_2-l_1)-MH}=\frac{f_m-f_o}{f_m-f_1}$$

i.e. $$(f_m-f_1)\ MH=(f_m-f_o)\{(l_2-l_1)-MH\}$$

i.e. $$MH\{(f_m-f_1)+(f_m-f_o)\}=(l_2-l_1)(f_m-f_o)$$

i.e. $$MH=(l_2-l_1)\frac{f_m-f_o}{2f_m-f_o-f_1}$$

Obtain the missing frequencies and the mode.

Class interval	Frequency
100-110	4
110-120	7
120-130	15
130-140	–
140-150	40
150-160	–
160-170	16
170-180	10
180-190	6
190-200	3
Total	**150**

Solution: Let the missing frequencies be denoted by f_1 and f_2. Then we have the frequency distribution given by :

130-140	f_1	$26+f_1$	24
140-150	40	$66+f_1$	40
150-160	f_2	$66+f_1+f_2$	25
160-170	16	$82+f_1+f_2$	16
170-180	10	$92+f_1+f_2$	10
180-190	6	$98+f_1+f_2$	6
190-200	3	$101+f_1+f_2$	3
Total	**150**		**150**

Since the total frequency is 150 and the sum of the given frequencies is 101, we have,

$$101+f_1+f_2=150$$

i.e. $f_2=150-101-f_1=49-f_1$

The median is 146.25 which lies in the class interval 140-150

We know, Median $=l_1+(l_2-l_1)(\frac{N}{2}-F)/f$

i.e. $146.25=140+10\,(75-26-f_1)\,/40$

$$=140+\frac{49-f_1}{4}$$

∴ $\frac{49-f_1}{4}=146.25-140=6.25$

i.e., $49-f_1=6.25\times4=25$

i.e., $f_1=49-25=24$

$$= 140 + 10 \times \frac{16}{31}$$

$$= 140 + 5.16$$

$$= 145.16$$

The symbols used in this formula have their usual meanings.

Note 1. For a frequency distribution in which all the observations have the same frequency, there exists no mode.

Note 2. For a moderately asymmetrical frequency distribution, mode can be determined by using the 'empirical formula'

$$\text{Mode} = 3 \text{ median} - 2 \text{ mean} \qquad (4.10)$$

4.6.5. Merits and Demerits of Mode.

Merits:

(i) Mode is simple to understand and easy to calculate. In some cases it can be determined by mere inspection.

(ii) Mode can be determined for data having open end classification. Even for unequal class intervals it can be determined provided that the modal class and its two adjacent class intervals are of equal magnitude.

(iii) It is not affected by extreme values of the observations

(iv) It can be used to describe qualitative data like brand preferences of consumer products, media preference for advertisements etc.

(v) It can be located graphically.

(iv) It is affected, to a greater extent, by fluctuations of sampling in comparison with the arithmetic mean.

Uses:

Inspite of several demerits, mode is frequently used in the study relating to business, industry and marketing.

4.7. GEOMETRIC MEAN

Geometric mean of 'n' positive numbers is defined as the nth root of their product.

Let $X_1, X_2, \ldots X_n$ be 'n' positive numbers. The geometric mean of these n numbers, G is given by

$$G = (X_1, X_2, \ldots X_n)^{1/n} \quad (4.11)$$

Thus, the geometric mean of three numbers 2,9,12 is

$$(2 \times 9 \times 12)^{1/3} = (216)^{1/3} = 6$$

If the number of observations is large, computation of geometric mean becomes tedious and cumbersome. In such situations, the computation can be facilitated by use of logarithms.

Since,

$$G = (X_1, X_2, \ldots X_n)^{1/n}$$

taking logarithm of both sides, we have,

$$\log G = \frac{1}{n} \log (X_1 X_2 \ldots\ldots X_n)$$

$$= \frac{1}{n} (\log X_1 + \log X_2 + \ldots\ldots + \log X_n)$$

geometric mean G is given by

$$G = (X_1^{f_1} X_2^{f_2} .....X_n^{f_n})^{1/N}, \text{ where } N = \sum_{i=1}^{n} f_i \qquad (4.12)$$

Using logarithm we can write (4.12) as

$$\log G = \frac{1}{N}[f_1 \log X_1 + f_2 \log X_2 + ... + f_n \log X_n]$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \log X_i$$

$$\text{So,} \quad G = \text{Anti log } [\frac{1}{N}\sum_{i=1}^{n} f_i \log X_i] \qquad (4.12a)$$

For a grouped frequency distribution, the class marks of the class intervals are taken as the X values.

Example 4. 18. Calculate the geometric mean for the following frequency distribution.

Class interval	Frequency
0-250	8
250-500	15
500-750	24
750-1000	38
1000-1250	29
1250-1500	18
1500-1750	11
1750-2000	7
Total	**150**

750-1000	38	875	2.9420	111.7980
1000-1250	29	1125	3.0511	88.4819
1250-1500	18	1375	3.1383	56.4894
1500-1750	11	1625	3.2108	35.3188
1750-2000	7	1875	3.2730	22.9110
Total	**150**			**437.4861**

$$\log G = \frac{1}{150} \times 437.4861$$

$$= 2.916574$$

Thus, geometric mean, G = Anti log 2.916574 = 825.228

4.7.1 Algebraic properties of Geometric Mean.

(i) If the geometric mean of 'n' numbers is known and the values of (n -1) of the numbers are known, the value of the nth number can be found.

Let the geometric mean of the 'n' numbers ($X_1, X_2, ..X_n$) be denoted by G. Then the n th number is given by,

$$X_n = \frac{G^n}{X_1 X_2 .....X_{n-1}}$$

(ii) If G_1 and G_2 be the geometric means of two sets of data with n_1 and n_2 observations respectively, then the geometric mean G of the combined set of (n_1+n_2) observations is given by

So, $X_1 X_2 .....X_{n1} = G_1^{n1}$ and $Y_1 Y_2 .....Y_{n2} = G_2^{n2}$

If the two distribution are combined , then the geometric mean of the combined distribution G is given by,

$$G = (X_1 X_2 .....X_{n1} Y_1 Y_2 .....Y_{n2})^{\frac{1}{n1+n2}} = \left(G_1^{n1}.G_2^{n2}\right)^{\frac{1}{n1+n2}}$$

The above formula can be extended to several sc.s of data.

Example 4.19. The G. M. of 5 observations was found to be 12. Later, it was found that an observation 9 was wrongly taken as 19. Calculate the correct value of the G.M.

Solution: Incorrect G. M. of 5 numbers = 12

Incorrect product of 5 numbers = 12^5

Correct product of 5 numbers $= \frac{12^5 x9}{19}$

Correct G. M. = $\left(\frac{12^5 \times 9}{19}\right)^{\frac{1}{5}}$

$$= \text{Antilog}\left[\frac{1}{5}\{5\log 12 + \log 9 - \log 19\}\right]$$

$$= 10.33$$

Example 4.20: The price of a commodity increased by 5% in the first year, 3% in the second year, 6% in the third year and 4% each in the fourth and fifth years. Find the average increase in the price of the commodity during these five years.

$= \text{Antilog } 0.63$

$= 4.28$

Example 4.21. The GMs of two groups containing 8 and 7 observations are respectively 10.37 and 13.78. Find the G. M. of all the 15 numbers by pooling the two groups together

$$G = \left(G_1^{n1} G_2^{n2}\right)^{\frac{1}{n1+n2}} = \left(10.37^8 \times 13.78^7\right)^{\frac{1}{15}}$$

$$= \text{Antilog}\left[\frac{1}{15}(8\log 10.37 + 7\log 13.78)\right]$$

$$= \text{Antilog}\left[\frac{8 \times 1.0149 + 7 \times 1.1393}{15}\right]$$

$$= \text{Antilog} 1.0729$$

$$= 11.82$$

Merits and Demerits of Geometric mean

Merits:

1. it is rigidly defined and based on all the observations of the data.
2. It gives more weight to small items and less to large items of the data.
3. It is capable of further algebraic treatments.

Demerits:

1. It is difficult to understand and difficult of compute.
2. It can not be calculated for data having positive and negative values.
3. The geometric mean of a set of numbers with one number zero is zero.

3. It is most suitable when the smaller values are to be given more weightage compared to the larger values.

4.8. HARMONIC MEAN

Harmonic mean of a set of positive numbers is defined as the reciprocal of the arithmetic mean of the reciprocals of the numbers.

Let $X_1, X_2, \ldots X_n$ be n numbers. The harmonic mean, H is given by

$$H = \text{Reciprocal of } \frac{1}{n}\left(\frac{1}{X_1}+\frac{1}{X_2}+\ldots+\frac{1}{X_n}\right) \tag{4.13}$$

i.e.,

$$\frac{1}{H}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i} \tag{4.13a}$$

Thus, the harmonic mean of the numbers 5, 10,16 and 20 is

$$H = \text{Rec.}\frac{1}{4}\left(\frac{1}{5}+\frac{1}{10}+\frac{1}{16}+\frac{1}{20}\right)$$

$$= \text{Rec.}\frac{1}{4}(0.2+0.1+0.0625+0.05)$$

$$= \text{Rec }(0.4125/4)$$

$$= \text{Rec }(0.103125)$$

$$= 9.697$$

If the data is given in the form of a frequency distribution (X_i, f_i), $i = 1,2,3,\ldots, n$, the harmonic mean H, is given by, i.e $$\frac{1}{H} = \frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{X_i} \qquad (4.13c)$$

Example 4.22. Calculate the harmonic mean of the following frequency distribution.

X	10	12	16	20	25	32	Total
f	6	9	12	8	5	4	44

Solution:

X	f	1/X	f/X
10	6	0.1000	0.6000
12	9	0.0833	0.7497
16	12	0.0625	0.3500
20	8	0.0500	0.4000
25	5	0.0400	0.2000
32	4	0.0313	0.1252
Total	**44**		**2.4249**

Harmonic Mean, $$H = \left[\left(\frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{X_i}\right)\right]^{-1} = \left(\frac{2.4249}{44}\right)^{-1}$$

$$= \frac{44}{2.4249} = 18.145$$

The average speed of the train , $H=\left[\frac{1}{2}\left(\frac{1}{75}+\frac{1}{60}\right)\right]$

$$=\left(\frac{1}{2}\times\frac{4+5}{300}\right)^{-1}$$

$$=\frac{600}{9}$$

$$=66.67 \text{ kmph}$$

Note: The same result will be obtained by dividing the total distance travelled by the total time taken.

4.8.1. Merits and Demerits of Harmonic Mean.

Merits

1. It is rigidly defined and is based on all the observations of the data.
2. It is suitable for further algebraic treatments. For example, if H_1 and H_2 be the harmonic means of two series consisting of N_1 and N_2 observations, the harmonic mean of the combined series, H containing $(N_1 + N_2)$ observations is given by

$$\frac{1}{H}=\frac{1}{N_1+N_2}\left(\frac{N_1}{H_1}+\frac{N_2}{H_2}\right) \qquad (4.14)$$

3. It gives greater weightage to smaller values.
4. It is not much affected by one or more large values.
5. It is typically suitable when a variable is averaged over time or as an average of prices where total expenditure is constant.

factor, viz. average speed over time factor, average rate of increase in profits of a concern or average price at which an article is sold or purchased.

4.9. RELATIONSHIP AMONG ARITHMETIC MEAN, GEOMETRIC MEAN AND HARMONIC MEAN.

Let A, G and H denote respectively the arithmetic mean, geometric mean, and harmonic mean of a series of positive numbers. Then,

(i) $A \geq G \geq H$ (4.15)

The equality holds only when the numbers are all equal

(ii) For two positive numbers a and b, $AH = G^2$ (4.16)

Example 4.23. The arithmetic mean of two numbers is 127.5 and their geometric mean is 60. Find the (i) harmonic mean and (ii) the two numbers.

Solution:

(i) We know that $AH = G^2$

Here $A = 127.5$ and $G = 60$

So, $127.5\ H = 60^2$

i.e. $H = 3600/127.5 = 28.24$

(ii) Let a and b be the two numbers and $a > b$

So, $\frac{a+b}{2} = 127.5$, and $\sqrt{ab} = 60$

or, $(a + b) = 2 \times 127.5 = 255$ and,

$ab = 60^2 = 3600$ (1)

From (1), and (2)

$$a = \frac{1}{2}[(a-b)+(a+b)] = \frac{1}{2}[255+225] = 240$$

$$b = \frac{1}{2}[(a+b)-(a-b)] = \frac{1}{2}[255-225] = 15$$

$\therefore$ The two numbers are 240 and 15.

4.10. PARTITION VALUES:

The values which divide a series of observations arranged in order of their magnitudes in to a number of equal parts are called partition values or quantiles or fractiles. Thus, median is a partition value. A series can be divided into a number of equal parts. viz. four, five , six , ten , one hundred etc. The partition values of such parts are called quartiles , pentiles, hexiles, deciles , percentiles etc. The commonly used partition values are Quartiles, Deciles and Percentiles.

4.10.1. Quartiles:

The values which divide a given series of arrayed observations in to four equal parts are called quartiles. Thus there are three quartiles which are denoted by Q_1 , Q_2 and Q_3 in ascending order . Q_1 is called the lower quartile and Q_3, the upper quartile. Q_1 is the value of the observation of the series which exceeds the lower 25% of the observations and is exceeded by the upper 75% of the observations. Similarly, Q_2, the second quartile exceeds 50% of the observations and is exceeded by the larger 50% of the observations. Hence Q_2 coincides with the median. Q_3 is that value of the series which exceeds the lower 75% of the observed values and is exceeded by the upper 25% of the values of the series.

Like the median, quartiles do not exist for every N. But , they can be computed with some assumptions of continuity.

For grouped data,

$$Q_i = l_1 + (l_2 - l_1)\left(\frac{\frac{iN}{4} - F}{f}\right), \quad i = 1,2,3$$

where l_1, l_2 are the lower and upper boundaries of the class interval containing the partition value and f is the frequency of that class interval, F is the cum. frequency of the preceding class interval and N is the total frequency.

4.10.2. Deciles.

The values which divide a series of arrayed observed values of a data into ten equal parts are called deciles. Thus there are nine deciles denoted by D_1, D_2, ...D_9. D_1 is called the first decile and exceeds the lower 10% of the observations and is exceeded by the upper 90% of the observations etc. So, $D_5 = Q_2$. The values of the deciles are calculated by using the following formula.

For ungrouped data,

$D_j = j(N+1)/10$th observation, $j = 1,2,,\ldots\ldots.9$

where N is the total number of observations.

For grouped data,

$$D_j = l_1 + (l_2 - l_1)\left(\frac{\frac{jN}{10} - F}{f}\right), \quad j = 1,2,\ldots,9$$

are called percentiles. Thus, there are 99 percentiles. P_5 is the 5th percentile which exceeds the lower 5% of the arrayed observations and is exceeded by the upper 95% of the observations. etc. So. $D_1 = P_{10}$, $Q_1 = P_{25}$, $Q_2 = D_5 = P_{50}$ etc. The percentile values are calculated by using the following formula

For Ungrouped Data :

$$P_k = \left(\frac{k(N+1)}{100}\right) \text{th observation of the arrayed data, } k = 1,2,\ldots 99$$

where N is the total frequency.

For Grouped Data :

$$P_k = l_1 + (l_2 - l_1)\left(\frac{\frac{kN}{100} - F}{f}\right), \quad k = 1,2,\ldots,99$$

where, l_1, l_2 and f are respectively the lower boundary, the upper boundary and the frequency of the class interval containing the concerned percentile, F is the cumulative frequency of the preceding class interva and N is the total frequency.

Steps involved in finding partition values:

For Ungrouped Data:

1. Arrange the observations of the data either in ascending or in descending order of their magnitude and then find the less than or more than cumulative frequencies.
2. Find the location of the concerned partition value by using the necessary formula. For example, the location of Q_3 would refer to the 3(N+1)/4th item from the smallest

1. If the class intervals are inclusive, convert these to exclusive to determine class boundaries.
2. Compute the less than or more than cumulative frequencies for all the class intervals.
3. Locate the position of the desired partition value by using the formula. For example, to determine P_{57}, the location of the partition value would be determined from the cumulative frequency $\frac{57N}{100}$. Similarly, for D_9, use the cumulative frequency $\frac{9N}{10}$
4. Determine the class interval containing the partition value having a location at a position in 3 above.
5. Apply the appropriate linear interpolation formula to compute the desired partition value.

Example 4.24. For the following frequency distribution :

X	12	13	14	18	20	25	30	Total
f	2	8	15	27	18	11	9	90

obtain Q_1, D_7 and P_{58}.

Solution.

X	f	CF
12	2	2
13	8	10
14	15	25
18	27	52
20	18	70
25	11	81
30	9	90
Total	90	

$= 63.7\text{th observation} = 64\text{th observation}$

$= 20$

$$P_{58} = \frac{58(N+1)}{100}\text{th observation} = 52.78\text{th observation}$$

$= 18 + 0,78(20 - 18)$

$= 18 + 1.56$

$= 19.56$

Example 4.25. Calculate

(i) the lower and upper quartiles

(ii) the ninth decile, and

(iii) the 47th percentile

for the following frequency distribution of the marks of 160 students in a class.

Marks	No. of Students
10-19	10
20-29	15
30-39	26
40-49	38
50-59	31
60-69	19
70-79	12
80-89	6
90-99	3
Total	**160**

40-49	39.5-49.5	38	89
50-59	49.5-59.5	31	120
60-69	59.5-69.5	19	139
70-79	69.5-79.5	12	151
80-89	79.5-89.5	6	157
90-99	89.5-99.5	3	160
Total		**160**	

(i) To find Q_1 , the lower quartile, we calculate the serial number N/4 = 160/4 = 40

The cum. frequency 40 corresponds the class interval 29.5 – 39.5. Applying interpolation formula,

$$Q_1 = l_1 + (l_2 - l_1)\left(\frac{\frac{N}{4} - F}{f}\right), \text{ we have}$$

$$Q_1 = 29.5 + (39.5 - 29.5)\left(\frac{40 - 25}{26}\right)$$

$$= 29.5 + \frac{10 \times 15}{26}$$

$$= 29.5 + 5.769$$

$$= 35.3 \text{ marks appx.}$$

To find Q_3, the upper quartile, we determine $\frac{3N}{4} = \frac{3x160}{4} = 120$. The cum. freq.

$= 49.5 + 10$

$= 59.5$

(ii) To find D_9, the ninth decile, we calculate the serial number $\frac{9N}{10}$ which is equal to $\frac{9 \times 160}{10} = 144$.

The cum. freq. 144 corresponds to the class interval 69.5-79.5

Applying interpolation formula,

$$D_9 = l_1 + (l_2 - l_1)\left(\frac{\frac{9N}{10} - F}{f}\right)$$, we have,

$$D_9 = 69.5 + 10 \times \frac{144 - 139}{12}$$

$$= 69.5 + 10 \times \frac{5}{12}$$

$$= 69.5 + 4.167$$

$= 73.667$ marks $= 73.7$ marks appx.

(iii) 47th percentile P_{47} is determined by finding the serial number $\frac{47xN}{100} = 75.2$ which corresponds to the class interval 39.5-49.5.

Applying interpolation formula, we get,

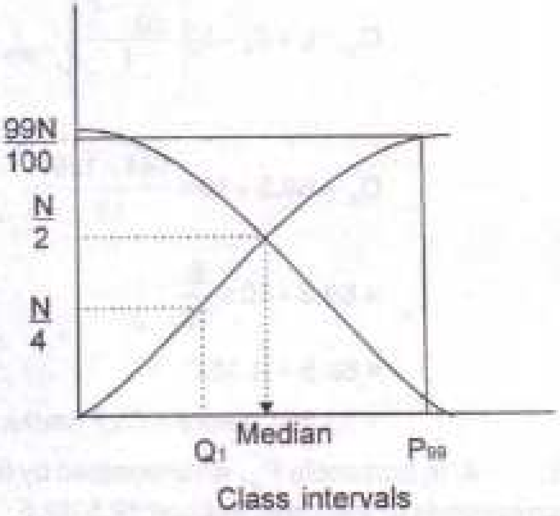
$$P_{47} = 39.5 + 10 \times \frac{75.2 - 51}{38}$$

$$= 39.5 + 6.368$$

4.11. DETERMINATION OF PARTITION VALUES GRAPHICALLY

All partition values like median , quartiles, deciles and percentiles of a grouped frequency distribution can be obtained graphically, either by drawing a less than or a more than cumulative frequency polygon as follows:

Median: To determine the median of a grouped frequency distribution consisting of N observed values, first a less than cumulative frequency polygon is drawn taking the values of the class intervals along the horizontal axis and the cumulative frequencies along the vertical axis. Since median corresponds to the (N/2) nd observation of the arrayed data, a point at a distance N/2 from the origin is marked along the vertical ordinate. From the point N/2, a line parallel to the horizontal axis is drawn which meets the cumulative frequency polygon at a point where from a perpendicular is drawn on the horizontal axis. The value corresponding to this point in the class interval gives the value of the median.



Median also can be determined by drawing both the less than and the more than cumulative frequency polygons. From the point of intersection of these two cumulative frequency polygons a perpendicular on the horizontal axis is drawn which gives the value of the abscissa as the median.

Other partition values : Quartiles, Deciles, Percentiles or any other partition values can be obtained graphically by drawing a cumulative frequency polygon and following similar steps as has been indicated for median.

Wages in Rs.	80-85	85-90	90-95	95-100	100-105	105-110	110-115	115-120	Total
No of workers	18	27	38	52	42	31	22	10	240

Also find (i) the number of workers with wages less than Rs. 102

(ii) the number of workers with wages between Rs. 93 and Rs. 112

(iii) the minimum wage of the highest paid 30 workers.

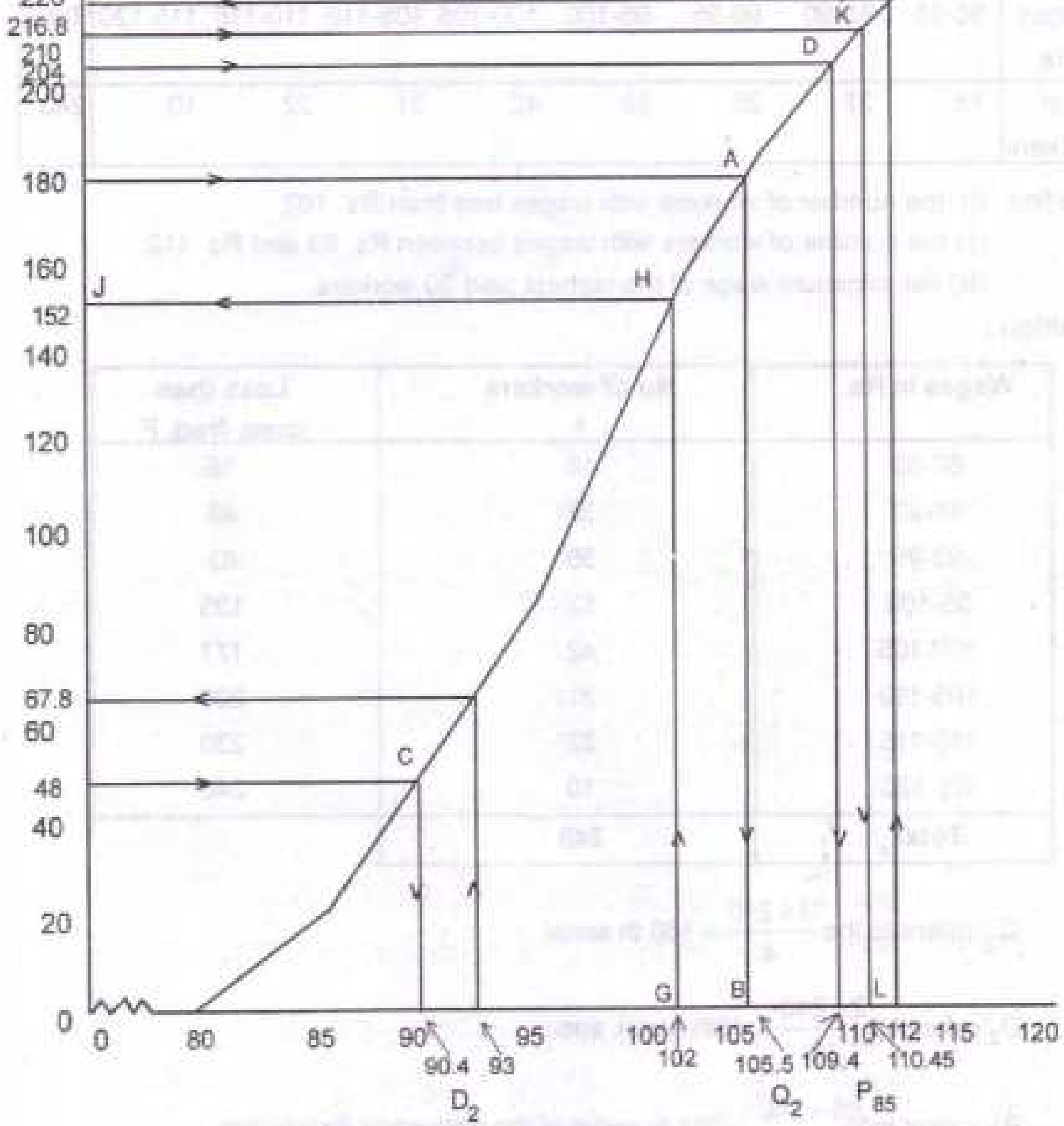
Solution :

Wages in Rs.	**No of workers** **f**	**Less than cum. freq. F**
80-85	18	18
85-90	27	45
90-95	38	83
95-100	52	135
100-105	42	177
105-110	31	208
110-115	22	230
115-120	10	240
Total	**240**	

Q_3 refers to the $\frac{3 \times 240}{4} = 180$ th serial

D_2 refers to $\frac{2 \times 240}{10} = 48$th serial, and

P_{85} refers to $\frac{85 \times 240}{100} = 204$ th serial of the frequency distribution.


220
216.8
210
204
200
180
160
152
140
120
100
80
67.8
60
48
40
20
0
K
D
A
H
J
C
G
B
L
0
80
85
90
90.4
93
95
100
102
105
105.5
109.4
110
112
110.45
115
120
D_2
Q_2
P_{85}

required partition values located on the abscissa.

To find the number of workers with wages less than Rs. 102, a perpendicular from G(102) is drawn to meet the cumulative frequency polygon at H. From H a parallel line is drawn to meet the Y-axis at J which gives the required number.

Similarly, the number of workers with wages less than Rs 112 and Rs 93 are computed. The difference is the number of workers with wages between Rs.93 and Rs. 112.

To find the minimum wage of the highest paid 30 workers, a line parallel to the horizontal axis at a distance 240 – 30 = 210 from the origin is drawn from the vertical axis meeting the cumulative frequency polygon at a point K where from a perpendicular KL is drawn on the horizontal axis to give the required minimum wage.

From the graph, we find

$Q_3 = 105.5$

$D_2 = 90.4$

$P_{85} = 109.4$

(i) The number of workers with wages less than Rs.102 is 152

(ii) The number of workers between Rs. 93 and Rs. 112 is 149

(iii) The minimum wage of the highest paid 30 workers is Rs.110.45

4.12. SELECTION OF AN AVERAGE.

From the discussion of various measures of central tendency it is clear that no single average is suitable for all purposes. Each of the averages has its own merits and demerits.

Arithmetic mean cannot be used for frequency distributions with open end class intervals. In such cases, median or mode are used. For qualitative data, median is suitable although in some cases of business activities like brand preference, mode is useful.

2. Whether the average is to be used in further computation and analysis.

3. The type of data available.

In case, the data represents extremely skewed distribution, arithmetic mean should not be used. In case the class intervals are not equal, avoid the mode and in case there are gaps around the middle of the distribution, avoid the median.

Median- Select median for open end grouped frequency distribution, J or U-shaped distributions or qualitative data.

Mode- Select mode for business activities or qualitative data involving preference or opinion.

Geometric mean - Select geometric mean in averaging the ratios, percentages, rates of increase or decrease.

Harmonic mean- Select harmonic mean as an average for quantities purchased or sold per unit, rate of speed per unit of time.

Arithmetic mean- Arithmetic mean is very commonly used average for most of the purposes. But the use should be avoided for open end distributions, extremely skewed distributions and unevenly spread distributions (concentration of data is larger or smaller at irregular points), for ratios and rates or for data having concentration of very large or very small observations.

EXCERCISES-4

1. What is an average ? Write the desirable properties of a good average.

2. Mention different types of averages and state why the arithmetic mean is the most commonly used average.

3. State important objectives of measures of central tendency.

4. State the properties of a good average. Examine these properties with reference to the Arithmetic Mean, the Geometric Mean and the Harmonic Mean. Give an example for each of these measures to be an appropriate measure of the average.

of them. Show how the geometric mean is capable of further mathematical treatments.

8. What are the requisites of a satisfactory average? In this light, compare the relative merits and demerits of three well known averages.

9. What are the chief measures of central tendency? Discuss their merits.

10. Show with the help of an example, that the

    (i) sum of the deviations about the arithmetic mean is zero.

    (ii) sum of the absolute deviations about the median is the least

    (iii) sum of the squares of deviations about the arithmetic mean is the least.

11. Three persons A, B and C were given to find the average of 5000 numbers in 10 equal groups, consisting of 500 units. They followed their own methods:

    A's method: He found the average of each of the 500 numbers separately and then found the average of these ten averages.

    B's Method: He found the averages of 2000 and 3000 numbers separately and then found the average of these two averages.

    C's Method: He united all the 500 numbers which were units and found the average of the remaining 4500 numbers. To this average he added 1.

    Are these methods correct? Give reasons.

12. Given below is the distribution of 140 candidates obtaining marks X or higher in a certain examination (all marks are given in whole numbers)

X	10	20	30	40	50	60	70	80	90	100
f	140	133	118	100	75	45	25	9	2	0

Calculate the mean, median and mode of the distribution.

[Hints. Form a frequency distribution with class intervals 10-19, 20-29, ... 90-99 and apply the formula for computation].

4	30	85

Find the mean of the distribution.

14. Define a weighted mean. If several sets of observations are combined to a single set, show that the mean of the combined set is a weighted mean of the means of the individual sets.

15. The weighted geometric mean of three numbers 72.5 , 320 and 125 is 200. The weights for the first and second numbers are 2 and 4 respectively. Find the weight of the third number.

16. Define the weighted arithmetic mean of a set of numbers. Show that it is unaffected if each of the weights are multiplied by a common factor.

17. The mean marks obtained in an examination by a group of 100 students was found to be 49.96. The mean of the marks obtained in the same examination by another group of 200 students was 52.32 . Find the mean marks obtained by both the groups of students taken together.

18. The mean marks of 300 students in the subject Statistics is 45. The mean of the top 100 of them is 70 and that of the last 100 students is 20. Find the mean of the remaining 100 students.

19. The mean weight of 150 students in a class is 60 kg. The mean weights of the boys and girls are respectively 70 kg and 55 kg. Find the number of boys and the number of girls in the class.

20. The average wage of 49 out of 50 employees in a firm is Rs. 100.00. The wage of the 50th employee is Rs. 97.50 more than the average wage of all the 50 employees. Find the average wage of all the employees of the firm.

Class	0-20	20-40	40-60	60-80	80-100	Total
Frequency	17	f_1	32	f_2	19	120

23. The mean marks of 80 students in a class was found to be 40. Later, it was discovered that two scores 26 and 54 were read as 62 and 45. Find the correct value of the mean.

24. The average salary of all the employees in a factory was Rs. 5000. The average salary of male and female employees taken separately were Rs.5200 and Rs. 4200 respectively. Find the percentage of male and female employees of the factory.

25. From the following data, calculate the percentage of workers getting wages (a) more than Rs. 104 (b)between Rs 82 and Rs. 118 . Also find Q_1 and Q_3

Wages (Rs) :	60 -70	70 - 80	80 -90	90 -100	100 -110	110 -120	120 -130	130 140
No. of workers:	20	45	85	160	70	55	35	30

Hint : (a) No. of workers getting more than Rs. 104 $= \frac{110-104}{10} \times 70 + 55 + 35 + 30$

$= 162$

Percentage of workers getting more than Rs.104 is $\frac{162}{500} \times 100 = 32.4\%$.

(b) Percentage of workers between Rs.82 and Rs.118

$$= \left[\frac{90-82}{10} \times 85 + \frac{118-110}{10} \times 55 + 160 + 70\right] \times \frac{100}{500} = 68.4\%$$

26. For the two frequency distributions given below, the means of I and II distributions are respectively 25.4 and 32.5 . Find the values of x and y.

The fish are classified according to their weights and are given in the following frequency distribution. It was known that the median weights in the before and after drying are 20.83 oz and 17.35 oz respectively. Some frequencies a and b in before drying and x and y in after drying are missing. It is known that a = x/3 and b = y/2. Find the missing frequencies

Weights(oz)	0 - 5	5 -10	10 -15	15 - 20	20 - 25	25 - 30
Before drying frequency :	a	b	11	52	75	22
After drying frequency :	x	y	40	50	30	28

28. From the following table showing the weights distribution of a commodity, determine (a) the mean (b) the median (c) the mode (d) the limits for weights of the middle 50% of the commodity, (e) the percentage of weights between 75 oz and 125 oz, (f) the percentage of commodity having weights more than 150 oz and (g) the percentage of the weights less than 100 oz.

Weights in oz	20-40	40-60	60-80	80-100	100-120	120-140	140-160	160-180	180-200
Number	8	12	20	30	40	35	10	7	5

29. From the following table showing the frequency distribution of marks of 65 students in a class, calculate the

(i) upper and lower quartiles

(ii) number of students who secured marks more than 17

(iii) number of students securing marks between 10 and 15.

Age in years	Under 25	25-29	30-34	34-44	45-54	55-64	65-74	Above 74	Total
No. of family	23	41	53	106	97	68	44	18	450

31. Define Geometric mean and Harmonic mean and explain their uses in statistical analysis.

32. A goods train runs 25 kms at a speed of 30 kmph, another 50 kms at the speed of 40 kmph, then due to repairs of the track, travels for 6 minutes at a speed of 10 kmph and finally covers the remaining 24 kms at a speed of 24 kmph. What is the average speed of the train in kmph?

33. In the following frequency distribution of marks of 100 students in a class, two frequencies f_1 and f_2 are missing. But the median and the mode are respectively 25 and 24. Calculate the missing frequencies and the mean

Marks	0-10	10-20	20-30	30-40	40-50
No. of Students	14	f_1	27	f_2	15

34. The frequencies of the numbers 3.2, 5.8, 7.9 and 4.5 are respectively x, x +2, x – 3 and x +6. If their arithmetic mean is 4.876, find x.

35. If G_1 is the geometric mean of N x's and G_2 is the geometric mean of N y's, show that G, the geometric mean of all the 2N values is given by, $G^2 = G_1G_2$

36. A variate takes values a, ar, ar^2, ..., ar^{n-1}each with unit frequency. If A is the arithmetic mean, G is the geometric mean and H is the harmonic mean, show that

$$A = \frac{a(1-r^n)}{n(1-r)},\ G = ar^{(n-1)/2},\ H = \frac{an(1-r)r^{n-1}}{1-r^n}$$

mean, median and mode of the new distribution are given in terms of those of the first distribution by the same transformation.

39. In a frequency distribution, the upper boundary of each class interval has constant ratio to the lower boundary. Show that the geometric mean G may be expressed by the formula:

$$\log G = X_0 + \frac{C}{N}\sum_i f_i(i-1)$$

Where X_0 is the logarithm of the mid value of the first interval and C is the logarithm of the ratio between upper and lower boundaries, and other symbols have their usual meanings.

40. Match each of the following item of groups A with the correct item of group B

Group (A)	Group (B)
(a) Arithmetic mean	(i) $l_1 + (l_2 - l_1)\left(\frac{N}{2} - F\right)/f$
(b) Geometric mean	(ii) $l_1 + (l_2 - l_1) f_2/(f_1 + f_2)$
(c) Harmonic mean	(iii) $\sum f X / \sum f$
(d) Median	(iv) $(X_1 X_2 ... X_n)^{1/n}$
(e) Mode	(v) $\left[\frac{1}{n}\left(\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_n}\right)\right]^{-1}$
	(vi) $l_1 + (l_2 - l_1)(f_m - f_0)/(2f_m - f_0 - f_1)$

(v) Percapita income in several countries.

(vi) Sale of shirts of sizes 38,39, 40,41, 42, 43,44

(vii) Marks obtained 8,10,12,4,7,11 and X, (X >12)

42. Pick up the correct answer

(a) For most uni-modal frequency distributions

(i) Mean lies between Median and Mode

(ii) Mode lies between Mean and Median

(iii) Median lies between Mean and Mode

(iv) Mode is the greatest value.

(b) The commonly used empirical relation between mean, median and mode to compute mode is

(i) Mean – median =3(mean-mode)

(ii) Mode = 3 mean – 2 median

(ii) Mode = 3 median – 2 mean

(iv) 3 median =2 mean – mode

(c) The geometric mean of 2, 9, 12, is

(i) 6 (ii) 9 (iii) 7,67, (iv) 7

(d) The strength of seven colleges in a state are 385,1748, 1343, 1935, 786, 2874 and 2108. The median strength is

(i) 1935 (ii)1748 (iii) 1597 (iv) None of these

(i) 100 (ii) 1 (iii) 0 (iv) None of these

(g) The most stable measure of central tendency is

(i) the mean (ii) the median (iii) the mode (iv) none of these

(h) The mean of 7 observations is 10 and the mean of 3 other observations is 5. The mean of 10 observations taken together is

(i) 5 (ii) 15 (iii) 7.5 (iv) 8.5

(i) A variable X takes values 1, 2, 3, ..., n ; each with unit frequency. The mean of X is

(i) n(n+1)/2 (ii) n/2 (iii) (n+1)/2 (iv) none of these

(j) The point of intersection of the less than and more than cumulative frequency polygons corresponds to

(i) the mean (ii) the median (iii) the mode (iv) the geometric mean

(k) The arithmetic mean of 10 observations is 11. The arithmetic mean of 9 of these observations is 12. So the 10th observation is

(i) 10, (ii) 2 (iii) 6 (iv) 12

(l) Each of the 15 students secure 8 marks in an internal test. So the

(i) Mean < Median, (ii) Mean > Median (iii) Mode is 8 (iv) Mean = 8

(m) To find the average of a frequency distribution having open end class intervals, the most appropriate average is

(i) the arithmetic mean (ii) the median (iii) the mode (iv) none of these.

ANSWERS

Exercises - 4

11. A's method is correct, but not B's and C's. A's method is correct because he made groups of equal sizes. B's and C' methods are not correct because they took groups of different sizes.

12. Mean = 50.714, Median = 51.167, Mode = 52.8

13. 73.8 15. 3 17. 51.53 18. 45 19. Boys-50, Girls-100.

20. Rs. 101.99 21. 56.3 22. $f_1 = 28$, $f_2 = 24$ 23. 39.7

24. Male -80%, Female-20% 25. Q_1 = Rs. 87.05, Q_3 = Rs. 109.29

26. x = 3, y = 2, 27. a = 3, b = 6, x = 9, y = 12.

28. (a) 108.5 oz (b) 108.75 oz (c) 118.3 oz, (d) 81.25 oz, 129.3 oz

(e) 48 oz (f) 12 oz (g) 40 oz.

29. (i) $Q_1 = 5.58$, $Q_3 = 14.9$. (ii) No. of students having marks more than 17 is 10.6 ≈ 11 (iii) No. of students having marks between 10 and 15 is 15.6 ≈ 16.

30. Median = 45.2 years, Q_3 = 57.5 years, D_2 = 32.5 years, P_{35} = 38.32 years.

32. 31.41 kmph. 33. $f_1 = 25$, $f_2 = 24$, Mean = 25.1 marks. 33. x = 4

40. (a) (iii), (b) (iv), (c) (v), (d) (i), (e) (vi)

41. (i) Mean (ii) Median (iii) Mean (iv) Median (v) Mean (vi) Mode (vii) Median

(m) (ii), (n) (iii).

***

(iii) Geometric mean can not be found for a frequency distribution of heights.

(iv) Median is suitable average for open end classified data.

(b) Which of the following averages has no mathematical properties ?

(i) Arithmetic mean (ii) Geometric Mean

(iii) Mode (iv) Harmonic mean.

(c) Which of the following averages is not based on all the observations of the data ?

(i) Arithmetic mean (ii) Median

(iii) Weighted arithmetic mean (iv) Harmonic mean

(d) Which of the following measures can not be determined graphically ?

(i) Median (ii) Mode (iii) 3rd quartile (iv) Geometric mean

(e) Which of the following measures is not rigidly defined ?

(i) Arithmetic mean (ii) Mode

(iii) Geometric mean (iv) Weighted Geometric mean

(f) For averaging the ratios and percentages, which of the following averages would be appropriate ?

(i) Arithmetic mean (ii) Geometric mean (iii) Median (iv) Mode

(g) Which of the following is correct to compute mode when it is ill defined ?

(i) Mode = 3 mean – 2 median (ii) Mode = 3 median – 2 mean

(iii) Mode = 2 median – 3 mean (iv) Mode = 2 mean – 3 median

(h) If A is the arithmetic mean, G is the geometric mean and H is the harmonic mean of the marks of 20 students in a class, then which of the following is true ?

(i) $A > G > H$, (ii) $A \geq G > H$ (iii) $A > G \geq H$ (iv) $A \geq G \geq H$

workers would be equal to which of the following ?

(i) ₹. 350 (ii) ₹. 360 (iii) ₹. 380 (iv) ₹. 370

2. **(a) Fill in the blanks :**

(i) In a moderately asymmetrical distribution, median lies between _______ and _______ .

(ii) Arithmetic mean is not suitable to compute the average of a frequency distribution having _______ classes.

(iii) To find the average speed over a time factor_______ is used as an average.

(iv) The sum of the differences of the observations of a data from its arithmetic mean is _______ .

(v) To find the average of a frequency distribution having open ends _______ or _______can be used.

(b) Indicate True (T) or False (F) in each of the following :

(i) Median can be determined graphically.

(ii) For 10 positive numbers not all equal, the relationship between A.M, G.M and H.M is A.M = G.M = H.M.

(iii) Arithmetic mean is the best to find the average of qualitative data.

(iv) Harmonic mean has mathematical properties.

(v) To find the lowest mark of the best 30 % students of a class consisting of 200 students, P_{70} is suitable.

3. Write short answers for the following questions :

(a) Write two merits of median.

(b) State three desirable properties of a good average.

(f) Write two uses of median.

(g) State two algebraic properties of geometric mean.

(h) Write two applications of harmonic mean.

(i) Explain what is meant by measures of central tendency.

(j) State the empirical relationship between mean, median and mode and write when to determine mode by using this relationship.

ANSWERS

1. (a) (iv) (b) (iii) (c) (ii) (d) (iv) (e) (ii)
   (f) (ii) (g) (ii) (h) (iv) (i) (ii) (j) (ii)

2. (a) (i) Mean and Mode (ii) Open end (iii) Harmonic mean
   (iv) 0 (v) Median, Mode.

   (b) (i) T (ii) F (iii) F (iv) T (v) T

***

5.1 INTRODUCTION :

The measures of central value, called averages, has been discussed in the previous chapter. These are single figures located at or around the centre of the distribution and are supposed to describe the characteristics of the entire data. But the averages, alone, are not enough to provide a complete picture of the data when there exist disparities in the values of the observations of a distribution. For two or more distributions the averages may be equal but still the distributions may differ from one another in a number of ways. This is because of the fact that, the observations of the distributions differ widely. So, it is essential to study the variability of the observations of a distribution along with a central value. In other words, to identify a series of observations, the central value must be supported and supplemented by some other measure called measure of dispersion.

As an illustration, consider the following example. Suppose the marks of three students A, B and C in five different subjects are as follows :

Student	Marks					Total	Mean
A:	30	30	30	30	30	150	30
B:	32	28	29	29	32	150	30
C:	3	19	30	57	41	150	30

We observe that, all the three series have the same mean viz, 30. So, by considering the mean alone, one might be tempted to conclude that they are indentical. But, a close analysis of the series will reveal that the three series differ widely from one another. In case of A, the mean is 30 and it is equal to each of the observation of the data. So for A, the arithmetic mean of the marks is the proper representative of the observations. For student B, the arithmetic mean of the marks is 30 and the difference of the observations from the arithmetic mean are not wide. So here also we can conclude that the arithmetic does not describe the distribution adequately and completely. So, the measures of central tendency should be supported by some other measures. One of such measures is termed as measures of dispersion.

The literary meaning of dispersion is "Scatteredness". We study dispersion to get an idea about the spread of the observations of a group or of a frequency distribution. This is an indicator of homogeneity or heterogeneity of observations in a distribution. In our illustration, we observe that the series of marks of A is stationary and shows no variation. Series B is slightly dispersed while series C is more dispersed. A mere inspection also indicates that series B is less homogeneous than series A and is more homogeneous than series C. In other words, series C is more heterogeneous in comparison with A and B.

Thus, dispersion or scatteredness or variability helps us to study the compactness of the distribution about a central value. A measure of dispersion is designed to state the extent to which there exist differences between the individual observations and some central or average value of a series. In measuring variation we would be interested in the extent of variation or the degree of variation but not the direction in which the observations vary.

5.2 OBJECTIVES OF STUDYING DISPERSION :

Measures of dispersion are needed for the following four basic purposes. These are :

(i) to determine the reliability of an average.

(ii) to control the variability of the data from the central value.

(iii) to compare the variability of two or more series.

(iv) to facilitate the use of other statistical measures.

We give below a brief explanation of each of these purposes.

(ii) Another objective of measuring dispersion is to determine the nature and cause of variation so that the variation itself can be controlled. For example, in matters of health, variations in body temperature and blood pressure are the basic guides to diagnosis, so that treatment can be prescribed to control them. Similarly, in industrial production, measurement of dispersion plays a vital role to control the causes of variation

(iii) Measures of dispersion is a devise to be used ‹› compare two or more series with regard to their variability. A high degree of variation would mean little uniformity or consistency where as a low degree of variation leads to great uniformity or consistency.

(iv) Many powerful analytical tools in statistics, such as study of correlation and regression, testing of hypothesis, quality control etc are based on measures of variation.

5.3 REQUIREMENTS (DESIDERATA) OF AN IDEAL MEASURE OF DISPERSION :

A good measure of dispersion should possess the following properties.

(i) It should be rigidly defined, simple to understand and easy to calculate.

(ii) It should be based on each and every item of the distribution.

(iii) It should be amenable to further mathematical treatments.

(iv) It should not be unduly affected by extreme items.

(v) It should have sampling stability.

5.4. ABSOLUTE AND RELATIVE MEASURES OF DISPERSION :

Measures of dispersion may be either absolute or relative. The measures of dispersion expressed in terms of the original units of the observations of a series such as independent of the unit of measurement of the data and is obtained as ratio and percentage.

5.5 TYPES OF MEASURES OF DISPERSION :

The various measures of dispersion are :

(i) Range

(ii) Inter quartile Range and Quartile Deviation

(iii) Mean Deviation

(iv) Standard Deviation

(v) Lorenz Curve

(i) and (ii) are called positional measures as they depend upon the values of the data at particular positions and (v) is called a graphical method of studying dispersion.

5.6 RANGE :

It is the simplest method of measuring dispersion. Range is defined as the difference between the value of the largest and the smallest item of a distribution. Symbolically,

$$\text{Range} = L - S \qquad (5.1)$$

where L = Largest item and S = Smallest item.

In case of a grouped frequency distribution, either of a discrete or of a continuous variable, range is defined as the difference between the upper boundary of the largest class and the lower boundary of the smallest class.

5.6.1 Absolute and Relative Measures of Range :

Range is an absolute measure of dispersion which depends upon the units of measurement of the characteristic under taken. So, it can be used to compare the variability of two or more distributions expressed in the same units of measurement. In case, of distributions having different units of measurements a relative measure of range, free

Merits :

(i) Among all the measures of dispersion, range is simplest, not only to understand, but also to calculate.

(ii) It is rigidly defined.

Demerits :

(i) It is not based on each and every item of the series.

(ii) It is very much affected by fluctuations of sampling.

(iii) It cannot be computed in case of open end distributions.

(iv) It is not suitable for further mathematical treatments.

5.6.3 Uses of Range :

Despite the above limitations, range has applications in a number of fields where the variation among the observations of the data is supposed to be small. We state below some of the fields of its applications.

(i) It is used for weather forecasts by meteorological department because, the general public may be inquisitive to know the limits within which the temperature is likely to vary in a particular day.

(ii) It is also used in studying variations in the prices of stocks and shares and for other commodities which are subject to price changes from time to time.

(iii) It is used in industry to check the quality of product, without making cent percent inspection, by constructing the Control Charts for Range.

Solution :

Since the frequency distribution is given in the form of inclusive classification, we make it exclusive classification to know the class boundaries.

Age in years	No of persons
14.5-19.5	15
19.5-24.5	27
24.5-29.5	35
29.5-34.5	13
34.5-39.5	10

Here the largest observation, L = 39.5 and the smallest observation, S = 14.5

$$\text{Coeffificent of Range} = \frac{L-S}{L+S} = \frac{39.5-14.5}{39.5+14.5} = \frac{25}{54} = 0.46$$

5.7 INTER QUARTILE RANGE AND QUARTILE DEVIATION :

Inter quartile range represents the difference between central 50% of the items i.e. the difference between the third quartile and the first quartile. Symbolically :

Inter Quartile Range = $Q_3 - Q_1$

Semi -Inter Quartile Range or Quartile Deviation is obtained by dividing Inter Quartile Range by 2.

Thus, Semi - Inter Quartile Range or Quartile Deviation (Q.D) is given by,

$$Q.D = \frac{Q_3 - Q_1}{2}$$

Quartile Deviation, being an absolute measure of dispersion, is unsuitable for comparison of variability. So to compare the variability of two or more distributions, a relative measure is computed. This relative measure is called Coefficient of Quartile Deviation.

1. It is quite easy to understand and compute.
2. It is computed by using the central 50 percent of the observations.
3. It is specially useful for distributions having open end classes.
4. It is not affected by the presence of extreme items in the data.

Demerits

1. Since it igonres the lower 25% and the upper 25% of the data, it cannot be regarded as a good measure of dispersion.
2. It is very much affected by fluctuations of sampling.
3. It is not suitable for further mathematical treatments.

Due to the above limitations, Q.D is not a reliable measure of variability.

Example 5.2

Calculate the quartile deviation of the monthly income of 7 families given below.

Family	1	2	3	4	5	6	7
Income (in Rs.)	1140	1265	1370	1185	1600	1197	1315

Solution :

Arranging the income in ascending order of magnitude we have,

1140, 1185, 1197, 1265, 1315, 1370, 1600

Q_1 = the size of $\left(\frac{N+1}{4}\right)$th. item = size of $\left(\frac{7+1}{4}\right)$th. item = size of the 2nd item.

Thus Q_1 = 1185

Q_3 = the size of 3 $\left(\frac{N+1}{4}\right)$th. item = size of 3 $\left(\frac{7+1}{4}\right)$th. item.

= size of the 6th item = 1370.

Marks :	10	18	30	38	40	48
No of students :	8	12	4	9	11	7

Solution :

Marks (X)	Frequency (f)	Cumulative freq. (C.F)
10	8	8
18	12	20
30	4	24
38	9	33
40	11	44
48	7	51
Total	51	

Q_1 = Size of $\left(\frac{N+1}{4}\right)$ th. item = size of $\left(\frac{51+1}{4}\right)$th. item = size of 13th. item.

Thus Q_1 = 18 marks

Q_3 = the size of 3 $\left(\frac{N+1}{4}\right)$ th. item = size of 39th. item.

Thus Q_3 = 40

Coeff of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40-18}{40+18} = \frac{22}{58} = 0.379$

30-40	5	5
40-50	10	15
50-60	18	33
60-70	23	56
70-80	28	84
80-90	8	92
90-100	4	96
Total	96	

$$Q.D = \frac{Q_3 - Q_1}{2}$$

Q_1 = size of the $\left(\frac{N}{4}\right)$th item = size of the $\left(\frac{96}{4}\right)$th item = size of the 24th. item.

Hence Q_1 lies in the class interval 50 - 60.

So, $Q_1 = l_1 + \frac{l_2 - l_1}{f}\left(\frac{N}{4} - F\right)$

$$= 50 + \frac{60-50}{18}(24-15) = 50 + \frac{10}{18} \times 9 = 50 + 5 = 55$$

Again, Q_3 is the size of the $\left(\frac{3N}{4}\right)$th item = size of the 3 $\left(\frac{96}{4}\right)$th item = size of the 72nd item.

∴ Q_3 lies in the class inerval 70 - 80.

$$Q.D = \frac{}{2} = \frac{}{2} = \frac{}{2} = 10.357$$

$$\text{Coefficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{75.714 - 55}{75.714 + 55} = \frac{20.714}{130.714} = 0.158$$

5.8 MEAN DEVIATION

Range and quartile deviation being positional measures of dispersion, are not based on all the observations and also do not show variation of the observations of a data about an average. To study the composition of observations in a distribution, the deviations of the observations from an average should be taken. Mean deviation is one of such measures of dispersion which is obtained by taking the average of the difference between the items of a distribution from a measure of central tendency, ignoring the algebraic signs. Let $x_1, x_2, \ldots\ldots\ldots\ldots x_n$ be a set of n observations of a data. The mean deviation (M.D) about an average A is defined as

$$M.D = \frac{1}{n}\sum_{i=1}^{n}|x_i - A| \qquad (5.6)$$

where A is any one of the averages viz. mean ($\overline{x}$) median (M_d) or mode (M_o).

In case of an ungrouped frequency distribution or a grouped frequency distribution or a frequency distribution of a continuous variable, mean deviation about an average A is given by

$$M.D\ (\text{about } A) = \frac{1}{N}\sum_{i=1}^{n} f_i|x_i - A| \qquad (5.7)$$

$$\text{where, } N = \sum_{i=1}^{n} f_i$$

Thus, mean deviation (about mean $\overline{x}$) $= \frac{1}{N}\sum_{i=1}^{n} f_i|x_i - \overline{x}|$

accompanying the symbols and the summation signs. For example one may write

M.D (about the mean, $\bar{x}$) as

$$\frac{1}{N}\Sigma f|x-\bar{x}| \text{ in stead of } \frac{1}{N}\sum_{i=1}^{n} f_i|x_i-\bar{x}|$$

Remarks

1. The sum of the absolute deviations of a given set of observations is minimum if deviations are taken from the median. It implies that the M.D about the median is always less than the mean deviation from the mean or the mode.

2. Since mode is ill defined, in actual practice, mean deviation is computed about the mean or the median. But since mean has got a wide range of applications in statistics, mean deviation from mean is frequently computed. So, by mean deviation, one usually refers to mean deviation about the mean.

3. The reason of taking the absolute deviations in the computation of mean deviation is due to the property that the "Algebraic sum of the deviations of a given set of observations from its arithmetic mean is always zero."

5.8.1 Merits and Demerits of Mean Deviation

Merits

(i) Mean deviation is rigidly defined and easy to understand and calculate.

(ii) It is based on all the observations and is thus a better measure of dispersion than range and quartile deviation.

(iii) It is less affected by the values of extreme items than the standard deviation (to be discussed in the next section)

Demerits

(i) The most serious drawback of mean deviation is that, in its computation, we ignore the algebraic signs which is mathematically illogical.

can not be used in sociological studies.

5.8.2 Uses of mean deviation

It spite of several drawbacks of mean deviation, it is frequently used in the field of economic phenomena and business statistics because of its simplicity. It is worth mentioning that the National Bureau of Economic Research in its work on forecasting business cycle has found the mean deviation to be the most practical measure of dispersion in the studies.

5.8.3 Relative Measures of Mean Deviation

Mean Deviations, as discussed above, are absolute measures of dispersion and depend on the units of measurement of the original data. As such, dispersion of two or more groups consisting of different units can not be compared through mean deviations. For purpose of comparison, relative measures of mean deviations are computed. Such measures are called coefficient of mean deviations.

$$\text{Coefficient of M.D (about an average A)} = \frac{\text{M.D about A}}{A} \qquad (5.9)$$

$$\text{Coefficient of mean Deviation (about mean, } \overline{x}\text{)} = \frac{\text{M.D about } \overline{x}}{\overline{x}} \qquad (5.9a)$$

$$\text{Coefficient of mean Deviation (about Median, Md)} = \frac{\text{M.D about } M_d}{M_d} \qquad (5.9b)$$

Age in years	No. of children f	Mid value x	$d=\frac{x-12}{5}$	fd	$\lvert x-\overline{x}\rvert$	$f\lvert x-\overline{x}\rvert$
0 - 4	5	2	−2	−10	10.56	52.80
5 - 9	15	7	−1	−15	5.56	83.40
10 - 14	32	12	0	0	0.56	17.92
15 - 19	22	17	1	22	4.44	97.68
20 - 24	6	22	2	12	9.44	56.64
Total	N = 80			$\sum fd=9$		$\sum f\lvert x-\overline{x}\rvert=308.44$

$$\overline{x}=A+\frac{\sum fd}{N}\times h=12+\left(\frac{9}{80}\right)\times 5=12+\frac{9}{16}=12+0.56=12.56$$

Mean Deviation about the mean, $\text{M.D}=\frac{1}{N}\sum_{i=1}^{n}f_i\left|X_i-\overline{X}\right|=\frac{308.44}{80}=3.86$

Coefficient of Mean Deviation about the mean

$$=\frac{\text{M.D}}{\overline{x}}=\frac{3.86}{12.56}=0.3069=0.31(\text{Approx})$$

Example - 5.6

Calculate the mean deviation about the median for the following data.

Size	6	7	8	9	10	11
Frequency	4	8	15	13	7	3

6	4	04	2	8
7	8	12	1	8
8	15	27	0	0
9	13	40	1	13
10	7	47	2	14
11	3	50	3	09
Total	50			52

Mean Deviation about the median = $\frac{\sum f|x-M_d|}{N} = \frac{52}{50} = 1.4$

Example 5.7

Calculate the mean Deviation about the median for the age distribution of 45 persons of a locality given below :

Age in years	0-10	10 - 20	20 - 30	30 - 40	40 - 50	Total
No of Persons	4	8	12	16	5	45

20 - 30	12	24	25	3.75	45.00
30 - 40	16	40	35	6.25	100.00
40 - 50	5	45	45	16.25	81.25
Total	45				431.25

Median = size of the $\left(\frac{N}{2}\right)$nd item = size of the 22.5th item.

Hence, the median class is (20 – 30).

$$M_d = l_1 + \frac{l_2 - l_1}{f}\left(\frac{N}{2} - F\right) = 20 + \frac{10}{12}(22.5 - 12) = 20 + \frac{10}{12}(10.5)$$

$$= 20 + 8.75 = 28.75$$

Mean Deviation about the Median

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \mid (x_i - M_d) \mid = \frac{431.25}{45} = 9.58$$

Example - 5.8

Calculate the mean deviation about the mean and the mean deviation about the median for the following frequency distribution and show that the former is greater than the latter.

Class Interval	100-110	110-120	120-130	130-140	140-150	150-160	160-170
Frequency	6	10	14	8	6	4	2

110-120	10	16	115	−2	−20	13.6	136.0	11.43	114.30
120-130	14	30	125	−1	−14	3.6	50.4	1.43	20.02
130-140	8	38	135	0	0	6.4	51.2	8.57	68.56
140-150	6	44	145	1	6	16.4	98.4	18.57	111.42
150-160	4	48	155	2	8	26.4	105.6	28.57	114.28
160-170	2	50	165	3	6	36.4	72.8	38.57	77.14
Total	50	–	–	–	−32	–	656.0	–	634.30

Mean, $\overline{x} = A + \frac{1}{N}\left(\sum fd\right)\ h = 135 + \frac{(-32)(10)}{50} = 135 - 6.4 = 128.6$

Median = $\frac{N}{2}$nd observation = 25th observation

25th observation lies in the class interval 120 -130

$$\therefore \quad \text{Median} = l_1 + \frac{(l_2 - l_1)\left(\frac{N}{2} - F\right)}{f}$$

$$= 120 + \frac{10}{14}(25 - 16) = 120 + \frac{90}{14} = 120 + 6.43 = 126.43$$

Mean Deviation about the mean, $\frac{1}{N}\sum f\,|\,x - \overline{x}\,| = \frac{656.0}{50} = 13.12$

Mean Deviation about the median, $\frac{1}{N}\sum f\,|\,x - Md\,| = \frac{634.3}{50} = 12.686$

Hence, M.D. about the Mean > M.D. about the Median.

observations, then the Standard Deviation (S.D), σ is given by,

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (5.10)$$

where $\bar{x}$ is the arithmetic mean of the given values.

In case of data represented in the form of a frequency distribution (x_i, f_i) $i = 1, 2, \ldots, n_i$ the S.D is given by

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \bar{x})^2} \qquad (5.11)$$

where, $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i$ and $N = \sum_{i=1}^{n} f_i$.

5.9.1 Merits and Demerits of Standard Deviation

Merits :

(i) Standard Deviation is a widely used measure of dispersion since it satisfies most of the important properties specified for an ideal measure of dispersion.

(ii) It is rigidly defined and based on all the observations of the data.

(iii) It is suitable for further mathematical treatments.

(iv) It is least affected by fluctuations of sampling.

Demerits :

(i) It is not readily comprehensible.

(ii) It gives greater weights to exterme values.

In spite of a few drawbacks, S. D is considered as the best and most powerful measure of dispersion in Statistical Theory.

and, for a frequency distribution, it is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{5.12a}$$

The symbols used here have the same meaning as has been indicated in sec. 5.9

5.9.3. Mean Square Deviation

The Mean Square Deviation is defined as the arithmetic mean of the squared deviations of the observations about any arbitrary value.

It is denoted by S^2 and is given by,

$$S^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^2 \tag{5.13}$$

where A is any arbitrary value

The positive square root of the Mean Square Deviation is termed as Root Mean Square Deviation and is given by :

$$S = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^2} \tag{5.14}$$

The relation between σ^2 and S^2 is : $S^2 \geq \sigma^2$

Proof : By definition,

$$S^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^2$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \overline{x} + \overline{x} - A)^2$$

$$[\ \because \frac{1}{N}\sum_{i=1} f_i(x_i-\bar{x})(\bar{x}-A) = (\bar{x}-A)\frac{1}{N}\sum_{i=1} f_i(x_i-\bar{x}) = (\bar{x}-A).0 = 0\ ]$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i(x_i-\bar{x})^2 + (\bar{x}-A)^2$$

So, $S^2 = \sigma^2$ +(a non-negative quantity)

$$\Rightarrow \quad S^2 \geq \sigma^2 \qquad (5.15)$$

So, we conclude that, the Mean Square Deviation is not less than the variance.

Further, if $\bar{x} = A$, $(\bar{x} - A)^2 = 0$. In this case,

$$S^2 = \sigma^2 \qquad (5.16)$$

In other words, S.D is the least possible value of Root Mean Square Deviations if the deviations are taken about the arithmetic mean.

5.9.4 Effect of Change of Origin:

Let x_i / f_i (i = 1, 2, ..., n) denote a frequency distribution where x_i's are the values of the variable and f_i's are their corresponding frequencies so that $\sum_{i=1}^{n} f_i = N$. Let $\bar{x}$ be the arithmetic mean and $\sigma^2 = \sigma_x^2$ be the variance.

$$\text{So,} \quad \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i-\bar{x})^2 \qquad (5.17)$$

Let $d_i = x_i - A$, where A is any arbitrary value

$\therefore x_i = A + d_i$

Putting the values of x_i and $\overline{x}$ in (5.17) we get,

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(A + d_i - A - \overline{d})^2$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i(d_i - \overline{d})^2$$

$$= \sigma_d^2$$

i.e. $\sigma_x = \sigma_d$.

This leads to the following important conclusion. The variance and consequently the S.D of a distribution is independent of change of the origin. By some writers this method of change of origin is called short-cut method.

Remark

By adding a constant to or substrating from each of the observations of a data, the variance and hence the S.D remains unaffected.

Effect of Change of Origin and Scale :

Let us examine to find what happens to the variance of an ungrouped or a grouped frequency distribution when, in addition to the change of origin, the scale is also changed.

Let σ_x^2 be the variance of a frequency distribution $x_i / f_i (i = 1, 2, \ldots, n)$, where $\sum_{i=1}^{n} f_i = N$ and $\overline{x}$ is the arithmetic mean.

Let $d_i = \frac{x_i - A}{h}$, where A is any arbitrary value and $h \neq 0$

So, $x_i = A + hd_i$ and $\overline{x} = A + h\overline{d}$ (by 5.17a)

$$= h^2 \cdot \frac{1}{N}\sum_{i=1}^{n} f_i \left(d_i - \bar{d}\right)^2$$

i.e. $\sigma_x{}^2 = h^2\ \sigma_d{}^2$ (5.18)

$\therefore\quad \sigma_x = |h| \cdot \sigma_d$ (5.18a)

(5.18) or (5.18a) do not contain A, the arbitrary origin but contain h, the scale.

So, we conclude that, the variance (or S.D) is independent of the origin, but is not independent of the scale.

5.9.5. An alternative form for calculation of variance

$$\sigma_x{}^2 = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - \bar{x}\right)^2 = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i{}^2 + \bar{x}^2 - 2x_i\,\bar{x}\right)$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i{}^2 + (\bar{x})^2\left(\frac{1}{N}\sum_{i=1}^{n} f_i\right) - 2\bar{x}\left(\frac{1}{N}\sum_{i=1}^{n} f_i\, x_i\right)$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i{}^2 + (\bar{x})^2 - 2(\bar{x})^2 = \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i{}^2 - (\bar{x})^2 \tag{5.19}$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i{}^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i\, x_i\right)^2 \qquad [5.19(a)]$$

When the values of the variable are integers and the mean $\bar{x}$ is fractional or involve decimals, (5.19) is a convenient form for the calculation of the variance. But if the values of x are large, computation of variance using (5.19) is tedious and time consuming.

In such cases, the variance of the distribution can be computed by using the following forms:

(5.21)

when the origin and scale both are changed i.e. $d_i = (x_i - A)/h$.

Example 5.9

Caculate the S.D for the following data.

x :	10	11	12	13	14
f :	3	12	18	12	3

Solution :

x	f	f x	fx²
10	3	30	300
11	12	132	1452
12	18	216	2592
13	12	156	2028
14	3	42	588
Total	48	576	6960

Using the form 5.19 we have,

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - (\bar{x})^2$$

$$\text{or, } \sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i x_i\right)^2} = \sqrt{\frac{6960}{48} - \left(\frac{576}{48}\right)^2}$$

$$= \sqrt{145 - 144} = \sqrt{1} = 1$$

12	18	0	0	0
13	12	1	12	12
14	3	2	6	12
Total	48		0	48

$$\sigma = \sqrt{\frac{1}{N}\Sigma fd^2 - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{48}{48} - 0} = \sqrt{1} = 1$$

Example 5.10

Compute the S.D of the following series

Class Interval	:	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25
frequency	:	7	12	19	10	2

Solution :

Calculation of S.D (by change of the origin and scale)

Class Interval	Mid Value X	Frequency f	$d = \frac{X - 12.5}{5}$	fd	fd^2
0 - 5	2.5	7	−2	−14	28
5 - 10	7.5	12	−1	−12	12
10 - 15	12.5	19	0	0	0
15 - 20	17.5	10	1	10	10
20 - 25	22.5	2	2	4	08
Total		50	0	−12	58

$= \sqrt{1.1024} \times 5 \quad = 1.05 \times 5 \quad = 5.25$

Example 5.11

Prove that, for any discrete series, S.D is not less than the mean deviation from the mean.

Solution :

We have to prove

S. D $\nless$ Mean Deviation about the mean.

$\Rightarrow$ S. D $\geq$ Mean Deviation about the mean.

We know, for a frequency distribution x_i / f_i $(i = 1, \ldots, n)$, $\sum_{i=1}^{n} f_i = N$

$$\text{S.D}, \sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^2} \quad \text{and}$$

$$\text{Mean Deviation about the mean} = \frac{1}{N}\sum_{i=1}^{n} f_i \,|\, x_i - \bar{x} \,|$$

To show $\sigma_x \nless$ Mean Deviation about mean is same as showing $\sigma_x \geq$ Mean Deviation about the mean i.e. $\sigma_x^2 \geq$ (Mean Deviation about the mean)^2

$$\text{i.e.} \quad \frac{1}{N}\sum_{i=1}^{n} f_i \,(x_i - \bar{x})^2 \geq \left[\frac{1}{N}\sum_{i=1}^{n} f_i \,\,|\, x_i - \bar{x} \,|\right]^2 \qquad (5.22)$$

Putting $|x_i - \bar{x}| = Z_i$, we have, $(x_i - \bar{x})^2 = Z_i^2$

Thus, (5.22) reduces to.

$$\Rightarrow \sigma_z^{\ 2} \geq 0$$

which is always true [$\sigma_z^{\ 2}$ being the variance of z is ≥ 0]

Hence the result.

5.9.6 Standard Deviation of Combined Series :

The S.D. of a number of series combined together can be obtained in terms of the mean, S.D. and the number of observations of the individual series.

Theorem :

Let there be two series with the following information :

		Series I	Series II
No. of observations	:	n_1	n_2
Mean	:	$\bar{x}_1$	$\bar{x}_2$
S.D.	:	σ_1	σ_2

The S.D, σ_{12} of the combined series I and II is given by

$$\sigma_{12} = \sqrt{\frac{1}{n_1 + n_2}\left[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)\right]} \tag{5.23}$$

where, $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$, and $\bar{x}$ is the mean of the combined series I and II.

Proof :

Let x_{1i}, $i = 1, 2, \ldots, n_1$ and x_{2j} $j = 1, 2, \ldots, n_2$ be the two series and

$$\bar{x}_1 = \sum_{i=1}^{n_1} \frac{x_{1i}}{n_1}, \quad \bar{x}_2 = \sum_{j=1}^{n} \frac{x_{2j}}{n_2},$$

$$\sigma_{12}^2 = \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(x_{1i}-\bar{x})^2 + \sum_{j=1}^{n_2}(x_{2j}-\bar{x})^2\right] \tag{5.24}$$

Now, $\sum_{i=1}^{n_1}(x_{1i}-\bar{x})^2 = \sum_{i=1}^{n_1}\left[(x_{1i}-\bar{x}_1)+(\bar{x}_1-\bar{x})\right]^2$

$$= \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + \sum_{i=1}^{n_1}(\bar{x}_1-\bar{x})^2 + 2\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)(\bar{x}_1-\bar{x})$$

$$= \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + n_1(\bar{x}_1-\bar{x})^2 + 2(\bar{x}_1-\bar{x})\sum_{i=1}^{n_1}(x_{1i}-\bar{x})$$

$$= \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + n_1(\bar{x}_1-\bar{x})^2. \quad \left[\because \ \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)=0\right]$$

$$= n_1\sigma_1^2 + n_1 d_1^2$$

Similarly, it can be shown that $\sum_{j=1}^{n_2}(x_{2j}-\bar{x})^2 = \sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2 + n_2(\bar{x}_2-\bar{x})^2$

$$= n_2\sigma_2^2 + n_2 d_2^2$$

So, $\sigma_{12}^2 = \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + n_1(\bar{x}_1-\bar{x})^2 + \sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2 + n_2(\bar{x}_2-\bar{x})^2\right]$

$$= \frac{1}{n_1+n_2}\left[n_1(\sigma_1^2+d_1^2)+n_2(\sigma_2^2+d_2^2)\right]$$

$$\therefore \quad \sigma_{12} = \sqrt{\frac{1}{n_1+n_2}\left[n_1(\sigma_1^2+d_1^2)+n_2(\sigma_2^2+d_2^2)\right]} \tag{5.25}$$

Note : The theorem can also be stated in another form as follows :

$$= \frac{n_1}{(n_1+n_2)^2}(n_1\bar{x}_1 + n_2\bar{x}_1 - n_1\bar{x}_1 - n_2\bar{x}_2)^2$$

$$+ \frac{n_2}{(n_1+n_2)^2}(n_1\bar{x}_2 + n_2\bar{x}_2 - n_1\bar{x}_1 - n_2\bar{x}_2)^2$$

$$= \frac{n_1 n_2^2}{(n_1+n_2)^2}(\bar{x}_1 - \bar{x}_2)^2 + \frac{n_1^2 n_2}{(n_1+n_2)^2}(\bar{x}_2 - \bar{x}_1)^2$$

$$= \frac{n_1 n_2}{(n_1+n_2)^2}(\bar{x}_1 - \bar{x}_2)^2 (n_2 + n_1) \qquad [\because (\bar{x}_2 - \bar{x}_1)^2 = (\bar{x}_1 - \bar{x}_2)^2]$$

$$= \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)^2 \tag{5.27}$$

Using 5.27 in 5.23 we get

$$\sigma_{12} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1+n_2} + \frac{n_1 n_2(\bar{x}_1 - \bar{x}_2)^2}{(n_1+n_2)^2}}$$

Note : The S.D. of k different groups combined to one can be found by using the formula given by

$$\sigma_{12\ldots k} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + \ldots + n_k(\sigma_k^2 + d_k^2)}{n_1 + n_2 + \ldots + n_k}}$$

where the symbols have the usual meanings.

Example 5.12

The means of two samples of sizes 35 and 65 are 80 and 70 and the standard deviations are 4 and 3 respectively. Obtain the standard deviation of the sample of size 100 obtained by combining the two samples.

$$\therefore \quad \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{(35 \times 80) + (65 \times 70)}{100} = \frac{2800 + 4550}{100} = \frac{7350}{100} = 73.5$$

So, $\quad d_1 = (\bar{x}_1 - \bar{x}) = (80 - 73.5) = 6.5$

and $\quad d_2 = (\bar{x}_2 - \bar{x}) = (70 - 73.5) = -3.5$

Putting these values in the formula for combined S.D given in (5.23) we get

$$\sigma_{12} = \sqrt{\frac{35(4^2 + 6.5^2) + 65(3^2 + 3.5^2)}{100}}$$

$$= \sqrt{\frac{35(16 + 42.25) + 65(9 + 12.25)}{100}}$$

$$= \sqrt{\frac{2038.75 + 1381.25}{100}} = \sqrt{\frac{3420}{100}} = \sqrt{34.2} = 5.848$$

Alternatively, using the formula (5.26), we have

$$\sigma_{12} = \sqrt{\frac{35(4)^2 + 65(3)^2}{35 + 65} + \frac{(35)(65)}{(35 + 65)^2}(80 - 70)^2}$$

$$= \sqrt{\frac{35 \times 16 + 65 \times 9}{100} + \frac{35 \times 65}{10000} \times 100}$$

$$= \sqrt{\frac{560 + 585}{100} + \frac{2275}{100}}$$

$$= \sqrt{11.45 + 22.75}$$

$$= 5.848 \text{ as before}$$

The S.D. $\sigma = \sqrt{\frac{\sum_{i=1}^{n} i^2}{n} - \left(\frac{\sum_{i=1}^{n} i}{n}\right)^2}$

$$= \sqrt{\frac{1^2 + 2^2 + \ldots + n^2}{n} - \left(\frac{1 + 2 + \ldots + n}{n}\right)^2}$$

$$= \sqrt{\frac{(n+1)(2n+1)}{6} - \left(\frac{(n+1)}{2}\right)^2} = \sqrt{\frac{(n+1)}{2}\left(\frac{(2n+1)}{3} - \frac{(n+1)}{2}\right)}$$

$$= \sqrt{\frac{(n+1)(n-1)}{12}} \quad = \sqrt{\frac{n^2-1}{12}}$$

5.10 COEFFICIENT OF VARIATION :

Standard deviation is an absolute measure of dispersion which depends upon the unit of measurement of the observations. A relative measure, called coefficient of S. D, is given by :

Coefficient of Standard Deviation $= \sigma / \overline{x}$ (5.28)

and is independent of units of measurements.

To compare the variability of two or more distributions, Karl Pearson devised a measure, called coefficient of variation. It is abbreviated as C.V. and is given by

$$C.V = 100 \times \sigma / \overline{x} \qquad (5.29)$$

A distribution with greater C.V is said to be less homogenous than the distribution with a smaller C.V. C.V is also used to indicate the consistency of distributions.

Who, between A and B, is more consistent worker ?

Solution :

We know, $\text{C.V} = \frac{\sigma}{\overline{x}} \times 100$

So, $\text{C.V (for worker A)} = \frac{8}{40} \times 100 = 20$

and $\text{C.V (for worker B)} = \frac{6}{42} \times 100 = \frac{100}{7} = 14.29$

Since the C.V of B is less than the C.V of A, worker B may be regarded as more consistent than worker A.

Example 5.15

Calculate the coefficient of variation for the following data.

Class Interval :	0-10	10-20	20-30	30-40	40-50
Frequency :	10	20	40	20	10

Solution :

Class Interval	Frequency f	Mid Value x	(x-25)/10 d	fd	fd^2
0 - 10	10	5	−2	−20	40
10 - 20	20	15	−1	−20	20
20 - 30	40	25	0	0	0
30 - 40	20	35	1	20	20
40 - 50	10	45	2	20	40
Total	100			0	120

Again, $\sigma = h\sqrt{\frac{\sum_{i=1}^{n} f_i d_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{N}\right)^2}$

$$= 10\sqrt{\frac{120}{100}} \quad = 10\sqrt{1.2} \quad = 10(1.095) \quad = 10.95$$

$$\therefore \quad C.V = \frac{10.95}{25} \times 100 = 43.8$$

Example 5.16

A variable takes values a, a+d, a+2d,..., a + 2nd, each with unit frequency. Find the mean diviation from the mean and the standard deviation. Also show that the mean deviation about the mean is less than the standard deviation.

Solution :

The mean $\bar{x}$, of the series is given by,

$$\bar{x} = \frac{a + (a + d) + (a + 2d) + ... + (a + 2nd)}{2n + 1}$$

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

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

$$= a + nd$$

	a + 2d	(n -2)d	$(n-2)^2 d^2$
	.....	:	:
	a+(n-1)d	d	d^2
	a+nd	0	0
	a+(n+1)d	d	d^2
	:	:	:
	a(2n-1)d	(n-1)d	$(n-1)^2 d^2$
	a+2nd	nd	n^2d^2
Total	(2n+1)a +n(2n+1)d	n(n+1)d	$\frac{n(n+1)(2n+1)d^2}{3}$

We find, $\sum_{i=1}^{2n+1} x_i = (2n+1)a + n(2n+1)d$

$$\sum_{i=1}^{2n+1} |x_i - \bar{x}| = n(n+1)d$$

and $\sum_{i=1}^{2n+1} (x_i - \bar{x})^2 = \frac{n(n+1)(2n+1)}{3} d^2$

$\therefore$ M.D about the mean $= \frac{1}{2n+1} \sum_{i=1}^{2n+1} |x_i - \bar{x}| = \frac{n(n+1)d}{2n+1}$

$$S.D = \sqrt{\frac{1}{2n+1} \sum_{i=1}^{2n+1} (x_i - \bar{x})^2} = \sqrt{\frac{n(n+1)(2n+1)}{3.(2n+1)} d^2}$$

$$= \sqrt{\frac{n(n+1)}{3}}\ d.$$

$$3 \qquad \left[\, 2n+1 \,\right]$$

i.e. if $(2n+1)^2 > 3n\,(n+1)$

i.e. if $4n^2+4n+1 > 3n^2+3n$

i.e. if $4n^2+4n+1-3n^2-3n > 0$

i.e. if $n^2+n+1 > 0$

i.e. if $\left(n+\frac{1}{2}\right)^2+\frac{3}{4} > 0$ (5.30)

But (5.30) is always true because

$\left(n+\frac{1}{2}\right)^2 > 0$ as n is a real number

∴ M.D. about the mean is less than the S. D.

5.11. LORENZ CURVE :

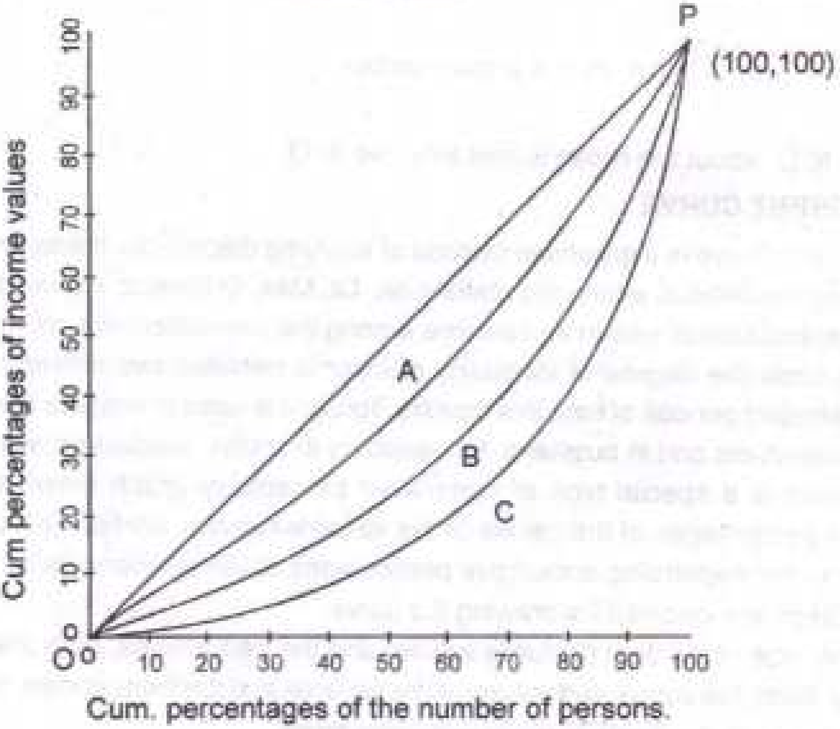
Lorenz Curve is a graphical method of studying dispersion. It was first suggested and used by the famous economic statistician, Dr. Max. O.Lorenz. He used the curve to show the distribution of wealth and income among the population of a country. The curve is used to show the degree of inequality of income between two different countries or between different periods of time in a country. Today, it is used to indicate the disparities in land, population etc and in business, for variability in profits, production, wages etc.

Lorenz Curve is a special type of cumulative percentage graph drawn by taking the cumulative percentages of the values of the variable (wealth, profits, turn overs) on one axis and the corresponding cumulative percentages of the frequencies on another. The following steps are involved for drawing the curve.

(i) The size of the item (variable values) and the frequencies, both are cumulated separately. Both, the cumulated values of the variable and the frequencies, are expressed as percentages on the basis of their respective totals.

(v) The percentages of the cumulated values of the variable (Y) and the percentages of the cumulated frequencies (X) for the given distribution i.e. the (X,Y) values are plotted on the graph as different points. Then these plotted points are joined by a smooth free hand curve. Obviously for any given distribution, this curve will never cross OP, the line of equal distribution. It will always lie below OP, unless the distribution is uniform. In the case of a uniform distribution, this plotted curve coincides with OP. If the curve is farther away from the line of equai distribution OP, greater variability is indicated. The greater the variability the greater is the distance of the curve from the line of equal distribution.

Lorenz Curve



area of concentration and the curves are called curve of concentration.

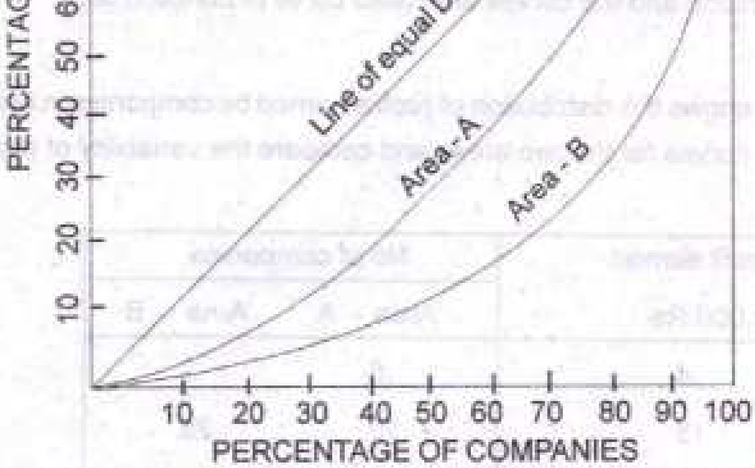
Example 5.17

The table below shows the distribution of profits earned by companies in two areas. A and B. Draw Lorenz curves for the two areas and compare the variability of profits in the two areas.

Profit earned	No of companies	
in 000 Rs	Area – A	A,na – B
4	6	2
15	10	25
50	12	22
75	15	18
90	14	13

Solution :

Profit earned Rs,000	Cumulative Profit	Cumulative Percentage	No of Comapny (Area – A)	Cumulative number (Area – A)	Cumulative Percentage (Area – A)	No of Comapny (Area – B)	Cumulative number (Area – B)	Cumulative Percentage (Area – B)
4	4	1.709	6	6	10.52	2	2	2.5
15	19	8.119	10	16	28.07	25	27	33.75
50	69	29.48	12	28	49.12	22	49	61.25
75	144	61.53	15	43	75.73	18	67	83.75
90	234	100.00	14	57	100.00	13	88	100.00



Since the distance of the Lorenz Curve for area B from the line of equal distribution is farther than for the Lorenz Curve for area A, we conclude that there exists a greater variability in area B than in area A so far as profit of the companies is concerned.

5.12. MOMENTS :

The term 'moment' is used in Meachanics as a measure of rotating effect of force F about some point, say A. It is denoted by the Greek alphabet μ (Mu) and is equal to the product of the force F and the perpendicular distance of A from O, the origin, Thus, $\mu = F.OA$. In Statistics, since we come across analogous quantities, we make use of the same symbol μ to denote the moment of a variable about a fixed value.

Let X be a discrete variable with the following frequency distribution :

x :	x_1	x_2	$....x_n$
f :	f_1	f_2	$....f_n$

So, $\sum_{i=1}^{n} f_i = N$ is the total number of observations and $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$ is the arithmetic mean of the distribution.

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^r, \quad r = 0, 1, 2, ... \tag{5.31}$$

and the rth moment about the origin as

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{N} f_i\, x_i^r \tag{5.32}$$

Putting r = 1 in (5.32) we have,

$$\mu_1' = \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i = \overline{x} \qquad \text{(Arithmetic mean)} \tag{5.33}$$

5.12.2. Central moments (or the moments about the arithmetic mean).

The r th central moment of a variable, denoted by μ_r, is given by

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \overline{x})^r \quad r = 0, 1, 2, ... \tag{5.34}$$

On putting r = 0 and r = 1 in (5.34) successively, we find,

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \overline{x})^0 = \frac{1}{N}\sum_{i=1}^{n} f_i = \frac{N}{N} = 1 \tag{5.35}$$

and $$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \overline{x}) = 0 \tag{5.36}$$

(being the algebraic sum of the deviations of the observations from the mean) Equations (5.35) and (5.36) are true for all distributions.

Again, putting r = 2 in (5.34), we get

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \overline{x})^2 = \sigma_x^2, \text{ the variance}$$

Thus we find, the second central moment of a variable is equal to its variance.

By the definition of the central moments,

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A + A - \overline{x})^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \left[(x_i - A) - (\overline{x} - A)\right]^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i (d_i - \mu_1')^r \tag{5.38}$$

where $x_i - A = d_i$ and $\overline{x} - A = \mu_1'$

Expanding the RHS of (5.38) by the Binomial Theorem we have,

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i \left[d_i^r - {}^rC_1\, d_i^{r-1}\mu_1' + {}^rC_2\, d_i^{r-2}(\mu_1')^2 - + ... + (-1)^r(\mu_1')^r\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\, d_i^r - {}^rC_1 \frac{1}{N}\sum_{i=1}^{n} f_i\, d_i^{r-1}(\mu_1') + {}^rC_2 \frac{1}{N}\sum_{i=1}^{n} f_i\, d_i^{r-2}(\mu_1')^2 + ... + (-1)^r \frac{1}{N}\sum_{i=1}^{n} f_i\, (\mu_1')^r$$

$$\mu_r' - {}^rC_1\, \mu_{r-1}'\mu_1' + {}^rC_2\, \mu_{r-2}'(\mu_1')^2 .... + (-1)^r(\mu_1')^r \tag{5.39}$$

$$\left[\because\ \frac{1}{N}\sum_{i=1}^{n} f_i d_i^r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^r = \mu_r'\right]$$

Putting r = 2, 3 and 4 successively in (5.39) we get,

$$\mu_4 = \mu_4' - {}^4C_1\,\mu_{4-1}'\mu_1' + {}^4C_2\,\mu_{4-2}'(\mu_1')^2 - {}^4C_3\,\mu_{4-3}'(\mu_1')^3 + {}^4C_4(\mu_1')^4$$

$$= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 4(\mu_1')^4 + (\mu_1')^4$$

$$= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 \tag{5.42}$$

5.12.4. Relation Between Raw moments and Central moments :

Let μ_r' be the r th moment about A.

$$\text{So, } \mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - A)^r = \frac{1}{N}\sum_{i=1}^{n} f_i\left[(x_i - \overline{x}) + (\overline{x} - A)\right]^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\left[(x_i - \overline{x}) + \mu_1'\right]^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \overline{x})^r + {}^rC_1\,\mu_1'\,\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \overline{x})^{r-1} + {}^rC_2\,(\mu_1')^2\,\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \overline{x})^{r-2}$$

$$+ \dots\dots\dots + {}^rC_{r-1}\,(\mu_1')^{r-1}\,\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \overline{x}) + \frac{1}{N}(\mu_1')^r.\sum_{i=1}^{n} f_i$$

$$\text{i.e. } \mu_r' = \mu_r + {}^rC_1\,\mu_{r-1}\mu_1' + {}^rC_2\,\mu_{r-2}(\mu_1')^2 + \dots + {}^rC_{r-2}\,\mu_2\,(\mu_1')^{r-2} + (\mu_1')^r \tag{5.44}$$

In particular, putting r = 2, 3 and 4 successively in (5.44) and simplifying, we get,

$$\mu_2' = \mu_2 + (\mu_1')^2 \tag{5.44a}$$

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + (\mu_1')^3 \tag{5.44b}$$

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\,(\mu_1')^2 + (\mu_1')^4 \tag{5.44c}$$

2. For computation of raw moments about the origin in terms of the central moments, μ_1' is to be replaced by $\bar{x}$, the first raw moment about the origin. Thus, we may write

$$\mu_2'(o) = \mu_2 + (\bar{x})^2,\ \mu_3'(o) = \mu_3 + 3\mu_2\,\bar{x} + (\bar{x})^3,$$

$$\mu_4'(o) = \mu_4 + 4\mu_3\,\bar{x} + 6\mu_2(\bar{x})^2 + (\bar{x})^4$$

5.12.5. Effect of Change of Origin and Scale on Central moments :

We know that, the r th central moment μ_r for X is given by

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \bar{x})^r \tag{5.46}$$

Let $u_i = \frac{x_i - a}{h}$, where 'a' is any arbitrary constant and $h \neq 0$

$\therefore \quad x_i = a + h\,u_i$

So, $\sum_{i=1}^{n} f_i\,x_i = a \sum_{i=1}^{n} f_i + h \sum_{i=1}^{n} f_i\,u_i$

i.e. $\frac{1}{N}\sum_{i=1}^{n} f_i\,x_i = a.\frac{1}{N}\sum_{i=1}^{n} f_i + h.\frac{1}{N}\sum_{i=1}^{n} f_i\,u_i$

i.e. $\bar{x} = a + h\,\bar{u}$ (5.47)

Putting (5.47) in (5.46), we have,

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i\,(a + hu_i - a - h\,\bar{u})^r$$

$$= h^r\,\frac{1}{N}\sum_{i=1}^{n} f_i\,(u_i - \bar{u})^r$$

$$= h^r\,\mu_r(u) \tag{5.48}$$

about the origin of the variable X is h' times the r th moment about the origin of U, i.e.

$$\frac{1}{N}\sum_{i=1}^{n} f_i x_i^r = h^r . \frac{1}{N}\sum_{i=1}^{n} f_i u_i^r .$$

5.12.6 Karl Pearson's Beta (β) and Gamma (γ) coefficients based on moments :

Karl Pearson defined the following coefficients based on the first four central moments.

$$\beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,3}} \qquad (5.49)$$

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} \qquad (5.50)$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^{\,3}}} \qquad (5.51)$$

$$\gamma_2 = \beta_2 - 3 \qquad (5.51)$$

These coefficients are pure numbers and are independent of units of measurements. These are used as measures of skewness and kurtosis which will be discussed later separately. For sake of convenience, γ_1 may be written as : $\gamma_1 = \frac{\mu_3}{\sigma^3}$

5.12.7 Sheppard's Correction for Moments :

In case of a grouped frequency distribution, mid-points of the class intervals are taken as the values of the variable for the purpose of calculation of moments. This concept of taking the middle points of the class intervals is based on the assumption that the frequencies are concentrated at the mid points of corresponding classes. For a symmetrical or a moderately asymmetrical distribution where the class intervals are not

(ii) the frequencies taper off to zero in both the directions,

the errors in the calculation of moments arise due to the consideration of the mid points as the values of the variable which can be corrected by using the following formulae. These are called Sheppards correction.

$$\mu_2 \text{ (Corrected)} = \mu_2 - \frac{h^2}{12}$$

$$\mu_3 \text{ (Corrected)} = \mu_3$$

$$\mu_4 \text{ (Corrected)} = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7}{240}h^4$$

where h is the magnitude of the class interval.

This is valid only for symmetrical or slightly asymmetrical continuous distributions and should not be applied in case of extermely skewed distributions.

Example 5.18.

Calculate the first four central moments of the following frequency distribution and hence find β_1 and β_2

x :	0	1	2	3	4	5	6
f :	5	15	20	35	10	10	5

4	10	1	10	10	10	10
5	10	2	20	40	80	160
6	5	3	15	45	135	405
Total	100		−20	220	−50	1240

Moments about the point 3 are :

$$\mu_1' = \frac{1}{N}\sum_{i=1}^{n} f_i d_i = -\frac{20}{100} = -0.2$$

$$\mu_2' = \frac{1}{N}\sum_{i=1}^{n} f_i d_i^2 = \frac{220}{100} = 2.2$$

$$\mu_3' = \frac{1}{N}\sum_{i=1}^{n} f_i d_i^3 = -\frac{50}{100} = -0.5$$

$$\mu_4' = \frac{1}{N}\sum_{i=1}^{n} f_i d_i^4 = \frac{1240}{100} = 12.4$$

So. $\mu_2 = \mu_2' - (\mu_1')^2 = 2.2 - 0.04 = 2.16$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$= -0.5 - 3(2.2)(-0.2) + 2(-0.2)^3$$

$$= -0.5 + 1.32 - 0.016 = 0.804$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

$$= 12.5 - 4(-0.5)(-0.2) + 6(2.2)(-0.2)^2 - 3(-0.2)^4$$

$$= 12.5 - 0.4 + 0.528 - 0.0048 = 12.6232$$

Example 5.19

Calculate the first four moments about the mean of the following frequency distribution.

Height (in inches) :	60 - 62	63 - 65	66 - 68	69 - 71	72 - 74
Frequency :	5	18	42	27	8

Solution :

Height inches	Freq f	Mid value x	$d=\frac{x-67}{3}$	fd	fd^2	fd^3	fd^4
60-62	5	61	−2	−10	20	−40	80
63-65	18	64	−1	−18	18	−18	18
66-68	42	67	0	0	0	0	0
69-71	27	70	1	27	27	27	27
72-74	8	73	2	16	32	64	128
Total	100			15	97	33	253

The raw moments about the aribitrary point A = 67 are given by :

$$\mu_1' = h\frac{\sum_{i=1}^{n} f_i d_i}{N} = 3\left(\frac{15}{100}\right) = 0.45$$

$$\mu_2' = h^2\frac{1}{N}\sum_{i=1}^{n} f_i d_i^2 = 3^2\left(\frac{97}{100}\right) = 8.73$$

$$\mu_3' = h^3\frac{1}{N}\sum_{i=1}^{n} f_i d_i^3 = 3^3\left(\frac{33}{100}\right) = 8.91$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 8.91 - 3(8.73)(0.45) + 2(0.45)^3$$

$$= 8.91 - 11.7855 + 0.18225 = -2.69325$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

$$= 204.93 - 4(8.9)(0.45) + 6(8.73)(.45)^2 - 3(0.45)^4$$

$$= 204.93 - 16.038 + 10.60695 - 0.1230 = 199.37595$$

Example 5.20

The first three moments of a distribution about the value 2 of the variable are 1, 16 and – 40 respectively. Find the first three moments about the mean and also the first three moments about the origin.

Solution :

We are given : $A = 2,\ \mu_1' = 1 \quad \mu_2' = 16 \quad$ and $\quad \mu_3' = -40$

Moments about the mean :

$$\mu_2 = \mu_2' - (\mu_1')^2 = 16 - 1^2 = 15$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$= -40 - 3(16)(1) + 2.(1)^3 = -40 - 48 + 2 = -86$$

$$\mu_1' \text{ (from the origin)} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \bar{x}$$

But we are given : $\mu_1'(2) = \frac{1}{N}\Sigma f_i(x_i - 2) = 1$ i.e. $\bar{x} - 2 = 1$

$\therefore \quad \bar{x} = 1 + 2 = 3$

Using equations (5.44) we have,

Prove that, for a discrete distribution, $\beta_2 \geq 1$

Solution :

Let the given distribution of a discrete variable X be given by (x_i / f_i), $i = 1, 2, \ldots, n$

where $N = \sum_{i=1}^{n} f_i$

By definition $\beta_2 = \frac{\mu_4}{\mu_2^2}$

where μ_4 and μ_2 are respectively the 4th and 2nd central moments of the variable X.

To show $\beta_2 \geq 1$ is same as showing

$$\frac{\mu_4}{\mu_2^2} \geq 1$$

i.e. $$\mu_4 \geq \mu_2^2$$

i.e. $$\frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^4 \geq \left[\frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2\right]^2$$

i.e. $$\frac{1}{N}\sum_{i=1}^{n} f_i z^2{}_i - \left(\frac{1}{N}\sum_{i=1}^{n} f_i z_i\right)^2 \geq 0, \text{ where } (x_i - \overline{x})^2 = z_i$$

But the LHS of the above inequality is the var (z)

So we find, Var $(z) \geq 0$

But. Var $(z) \geq 0$ is always true.

Hence, $\beta_2 \geq 1$

around a central value while measures of dispersion throw light on the scatter or spread of the observations of a data about some measure of central tendency. In reality, we may come across different distributions having same measures of central tendency and dispersion, yet differing very widely from one another in their compositions. i.e. in shapes and sizes. The following example will be instructive.

Consider the two different frequency distributions given below:

Frequency Distribution - A

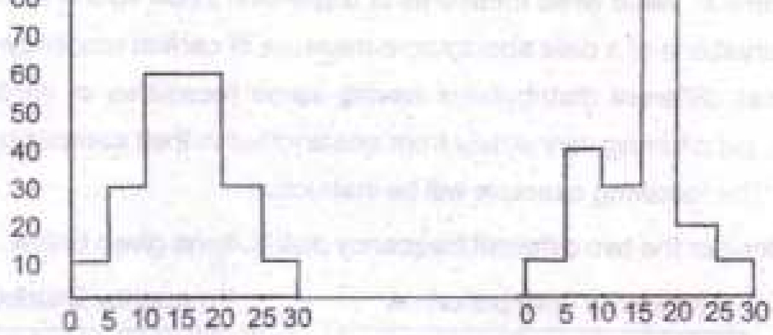
Class Interval	Frequency
0 - 5	10
5 - 10	30
10 - 15	60
15 - 20	60
20 - 25	30
25 - 30	10

Frequency Distribution - B

Class Interval	Frequency
0 - 5	10
5 - 10	40
10 - 15	30
15 - 20	90
20 - 25	20
25 - 30	10

On computation, we will find that , for the above two frequency distributions, the arithmetic means are 15 each and the standard deviations are 6 each. From the values of the arithmetic means and standard deviations, one may be tempted to infer that the two distributions are id . But, how erroneous such a conclusion would be evident from their Histograms given below.

We observe that the shape of the Histograms for the two distributions are not the same. The Histogram for the distribution A is symmetrical while that of B is asymmetrical. Further, the top of the Histogram for B has a larger peak than that of A.



Thus the central tendency, and dispersion are inadequate to describe a distribution completely. They must be studied along with other measures like Skewness and Kurtosis. Skewness helps in studying the shape while Kurtosis reflects on the flatness of the curve representing the distribution.

The term "skewness" means lack of symmetry or asymmetry. A distribution is said to be skewed or asymmetrical if its frequency curve is more stretched to one side than to the other. Skewness helps us to determine the nature and extent of the concentration of the observations, whether towards the higher or the lower values of the variable.

A distribution is said to be symmetrical if equal distances on either side of the central value have same frequencies and consequently, both the tails (left and right) of the curve are identical in shape and length.

5.13.2. Types of Skewness :

There can be two types of skewed distributions viz.

(a) Positively skewed distribution, and

(b) Negatively skewed distribution.

maximum and then decreases slowly. Here the mode has the smallest and the mean has the greatest value.

(b) Negatively Skewed Distribution

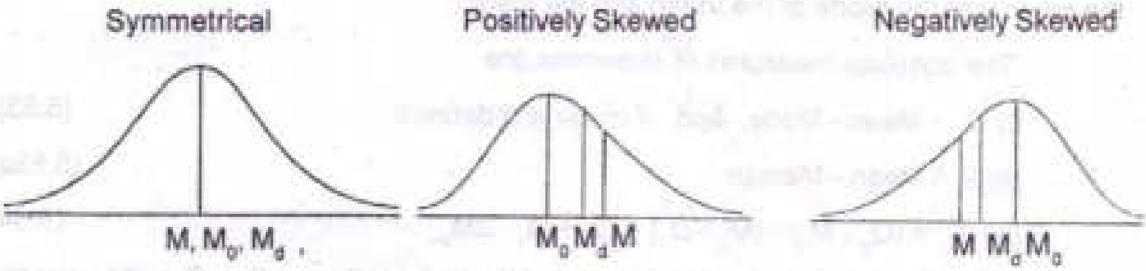
A frequency distribution is said to be negatively skewed if its frequency curve has

(i) a longer tail on the lefthand side than on the right hand side and

(ii) Mean < Median < Mode

In this case the frequencies increase slowly, reach a maximum and then decrease rapidly. Here the mean has the smallest and the mode has the greatest value.

We distinguish betwen the positively and negatively skewed curves with the help of the following diagram.



Mean = Median = Mode Mode < Median < Mean Mean < Median < Mode

In either of the cases of skewed distributions the median lies between the mean and the mode. In case, the mean and the mode are equal, the median will automatically coincide with the mean and the mode giving rise to a symmetrical curve.

5.13.3 Tests of Skewness

In order to examine the presence of skewness, the following tests may be applied.

(i) The values of the mean, the median and the mode do not coincide i.e

Mean ≠ Median ≠ Mode

(ii) The quartiles are equidistant from the median.

(iii) The frequencies are evenly distributed at points of equal distances from the mode.

(iv) The data, when plotted on a graph paper, gives a symmetrical form.

5.13.4. Measures of Skewness

Measures of skewness tell us the direction and extent of asymmetry of a series of observations. Thesa measures may be absolute or relative. We give below the various measures of skewness.

(a) Absolute measures of Skewness

Skewness can be measured in absolute terms by taking the difference between the mean and the mode or the mean and the median.

The absolute measures of skewness are :

(i) S_k = Mean – Mode, and if mode is ill defined, (5.53)

(ii) S_k = Mean – Median (5.53a)

(iii) $S_k = (Q_3 - M_d) - (M_d - Q_1) = Q_3 + Q_1 - 2M_d$ (5.54)

If the value of the mean is greater than the mode or the median, S_k will be positive indicating positive skewness and if the value of the mode is greater than the mean and the median, S_k will be negative, indicating negative skewness.

The above absolute measures of skewness are expressed along with the units of measurement of the observations. So, those are inconvenient for comparison of the skewess of two or more distributions with observations measured in different units. In such cases we take the help of relative measures of skewness, called the coefficient of skewness.

(iv) measure of Skewness based on moments.

(i) Karl Pearson's Coefficient of Skewness :

Karl Pearson's Coefficient of Skewness, also known as Pearsonian Coefficient of Skewness, Sk is given by

$$Sk = \frac{\text{Mean} - \text{Mode}}{\text{S.D}} = \frac{M - M_o}{\sigma} \tag{5.55}$$

When mode is ill defined and the distribution is modurately asymmetrical, Pearson's Coefficient of Skewness, Sk can be calculated by using the empirical relationship

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

$$\text{Thus,} \quad Sk = \frac{3(\text{Mean} - \text{Median})}{\text{S.D}} = \frac{3(M - M_d)}{\sigma} \tag{5.56}$$

Formula (5.56) is called Pearson's alternative formula for coefficient of skewness.

The coefficient of skewness given in (5,56) lies between the limits ± 3 but these limits are rarely attained in practice.

In case of a symmetrical distribution, the values of the mean, median and mode coincide. As a result, the coefficient of skewness becomes zero. For a positively skewed distribution, the coefficient of skewness is positive while for a negatively skewed distribution it is negative.

(ii) Bowley's Coefficient of Skewness :

A measure of skewness suggested by late prof. A.L. Bowley is based on quartiles. This measure is also known as Quartile Coefficient of Skewness.

Bowley's Coefficient of Skewness, S_k is given by

For two positive real numbers a and b, (i.e. a>0 and b>0), we know that

$$|a-b| \leq |a+b| \tag{5.58}$$

$$\Rightarrow \left|\frac{a-b}{a+b}\right| \leq 1$$

Since, for any distribution, we know that (Q_3-M_d) and (M_d-Q_1) are both non negative, taking $a = Q_3-M_d$ and $b = M_d-Q_1$ in (5.58), we get

$$\left|\frac{(Q_3-M_d)-(M_d-Q_1)}{(Q_3-M_d)+(M_d-Q_1)}\right| \leq 1$$

$$\Rightarrow \left|\frac{Q_3+Q_1-2M_d}{Q_3-Q_1}\right| \leq 1$$

So, $|S_k| \leq 1$

Thus, Bowley's coefficient of skewness lies between -1 and $+1$.

From the formula (5.57) it is obvious that,

$S_k = 1$ if $M_d-Q_1 = 0$ i.e. if $M_d = Q_1$ and

$S_k = -1$ if $Q_3-M_d = 0 \Rightarrow Q_3 = M_d$

Remark

Bowley's Coefficient of Skewness is specially used when (i) the distribution has open classes or unequal clas intervals and (ii) the mode is ill defined, but the quartiles and median can be computed. The draw back of this formula is that, it makes use of the central 50% of the data leaving the remaing 50% on the two extremities

Kelly's Coefficient of Skewness, $S(k) = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}$ (5.59)

or equivalently, $S(k) = \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$ (5.59a)

where P denotes the percentiles and D, the deciles.

This formula, although theoretically sounds better, is seldom used in practice. The formula due to Karl Pearson is popular for determination of coefficient of skewness.

(iv) Coefficient of Skewness based on Moments

There are two formulae, based on central moments, which are used as measures of coefficient of skewness viz.

(i) $\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3}$, and (5.60)

(ii) $S_k = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$ (5.60a)

where, $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ and $\beta_2 = \frac{\mu_4}{\mu_2^2}$

$\gamma_1 > 0$ or < 0 or $= 0$ according as $\mu_3 > 0$ or < 0 or $= 0$

Similarly, $S_k = 0$ if either $\beta_1 = 0$ or $\beta_2 = -3$

But, since $\beta_2 = \frac{\mu_4}{\mu_2^2}$ can not be negative, $S_k = 0$ only if $\beta_1 = 0$ i.e. $\mu_3 = 0$

value	freq.	value	freq.
6	4	25	15
12	6	30	10
18	8	42	4
22	18		

Solution :

value	freq	(x-22)	fd	fd²
x	f	d		
6	4	– 16	– 64	1024
12	6	– 10	– 60	600
18	8	–4	– 32	128
22	18	0	0	0
25	15	3	45	135
30	10	8	80	640
42	4	20	80	1600
Total	65		49	4127

Karl Pearson's Coefficient of Skewness

$$Sk = \frac{\text{Mean} - \text{Mode}}{\text{S.D}}$$

$$\text{Mean, } \bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i d_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{N}\right)} = \sqrt{\frac{4127}{65} - \left(\frac{49}{65}\right)^2}$$

$$= \sqrt{63.4923 - 0.5682} = \sqrt{62.9241} = 7.932$$

So, $\quad Sk = \dfrac{22.754 - 22}{7.932} = \dfrac{0.754}{7.932} = 0.095$

Example 5.23

Compute the Coefficient of Skewness for the following data.

Variable :	5-7	8-10	11-13	14-16	17-19
Freq :	7	12	19	10	2

Solution :

Variable	Mid Value(x)	freq f	$\frac{x-12}{3}$ d	fd	fd^2
4.5-7.5	6	7	−2	−14	28
7.5-10.5	9	12	−1	−12	12
10.5-13.5	12	19	0	0	0
13.5-16.5	15	10	1	10	10
16.5-19.5	18	2	2	4	8
Total		50		−12	58

Here, $A = 12$, $\sum_{i=1} f_i d_i = -12$, $h = 3$ and $N = 50$

So, the mean, $\bar{x} = 12 + 3\left(-\frac{12}{50}\right) = 12 - \frac{36}{50} = 12 - 0.72 = 11.28$

and, Mode $= l_1 + \frac{l_2 - l_1}{f} \frac{(f_m - f_0)}{(2f_m - f_0 - f_1)}$

Since the largest frequency 19 lies in the class interval 10.5 - 13.5, we have

$f_m = 19$, $f_0 = 12$ $f_1 = 10$, $l_1 = 10.5$ and $l_2 = 13.5$.

$\therefore$ Mode $= 10.5 + \frac{3(19-12)}{(2 \times 19 - 12 - 10)} = 10.5 + \frac{21}{16} = 10.5 + 1.3125 = 11.8125$

$$S.D = h\sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i d_i^2 - \left(\frac{\sum_{i=1}^{n} f_i d_i}{N}\right)^2} = 3\sqrt{\frac{58}{50} - \left(-\frac{12}{50}\right)^2}$$

$$= 3\sqrt{1.16 - 0.0576} = 3\sqrt{1.1024} = 3.15$$

$$\therefore \quad Sk = \frac{(\text{Mean} - \text{Mode})}{S.D} = \frac{11.28 - 11.81}{3.15}$$

$$= -\frac{0.53}{3.15} = -0.168 \simeq -0.17$$

Example 5.24

For the frequency distribution given below, calculate Bowely's Coefficient of Skewness.

Clas Interval	Frequency f	Cum freq F
9.5 - 19.5	5	5
19.5 - 29.5	9	14
29.5 - 39.5	14	28
39.5 - 49.5	20	48
49.5 - 59.5	25	73
59.5 - 69.5	15	88
69.5 - 79.5	8	96
79.5 - 89.5	4	100
Total	100	

Median = size of the $\left(\frac{N}{2}\right)$nd item = size of the $\left(\frac{100}{2}\right)$nd item

= size of the 50th item.

So, the median lies in the class interval 49.5 - 59.5

$$\text{Median } = l_1 + \frac{l_2 - l_1}{f}\left(\frac{N}{2} - F\right) = 49.5 + \frac{10}{25}(50 - 48)$$

$$= 49.5 + 0.8 = 50.3$$

Q_1 = size of the $\left(\frac{N}{4}\right)$th item = size of the $\left(\frac{100}{4}\right)$th item

$$= 29.5 + \frac{10}{14}(25-14) = 29.5 + \frac{110}{14} = 29.5 + 7.85 = 37.35$$

Q_3 = size of the $\left(\frac{3N}{4}\right)$th item i.e. size of the 75th item

The 75th item lies in the class interval 59.5 - 69.5

$$\text{So, } Q_3 = l_1 + \frac{l_2 - l_1}{f}\left(\frac{3N}{4} - F\right)$$

$$= 59.5 + \frac{10}{15}(75-73) = 59.5 + 1.3 = 60.8$$

Bowely's Coefficient of Skewness

$$S_k = \frac{(Q_3 + Q_1 - 2Q_2)}{Q_3 - Q_1} = \frac{60.8 + 37.35 - 2(50.3)}{60.8 - 37.35}$$

$$= \frac{98.15 - 100.6}{23.45} = \frac{-2.45}{23.45} = -0.1044$$

5.13.5. Kurtosis

The dictionary meaning of kurtosis is bulginess or the relative degree of sharpness of the peak of a frequency distribution curve. In Statistics, the term kurtosis refers to the degree of peakedness or flat toppedness of a frequency curve relative to the normal distribution curve.

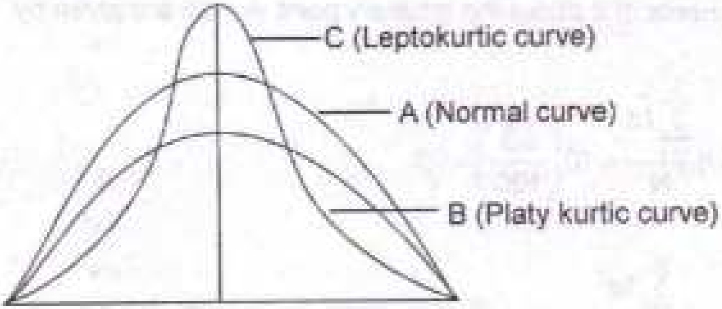
A measure of kurtosis tells us the extent to which a distribution has frequency curve having more peak or more flat top than the normal curve.

Measures of central tendency, dispersion and skewness, we have studied so far, do not reflect on the peakedness or the flatness of the top of a frequency curve. We

Freq. of A : f_A	4	8	16	32	50	32	16	8	4
Freq. of B : f_B	2	5	8	15	30	15	8	5	2
Freq. of C : f_C	7	11	22	40	67	40	22	11	7

From actual computation we will find that mean, median and mode of the three distributions are all equal to 72.5, the range are all equal to 45 and the S.Ds are all approximately equal to 1.66 indicating absence of skewness. But still on drawing the frequency curves of the three distributions we can mark the difference in their shapes. The shape of the curve for distribution A is bell shaped. It is called a normal curve. The shape of the curve for distribution B is more flat topped while that of C is more peaked topped than the normal curve. This difference in the shape of the curves is due to the presence of kurtosis.

Kurtic curves are of three types viz. Meso kurtic, Lepto kurtic and Platy kurtic. A bell-shaped curve is called Meso kurtic. It is also called normal curve. It is neither more flat topped nor more peak topped. A more peak topped curve than the meso kurtic curve is called Lepto kurtic curve and a more flat topped curve than the normal (meso kurtic) curve is platy kurtic curve, as has been indicateu in the following diagram.



< 0 the distribution is Platykurtic.

Example 5.25 :

Find the kurtosis for the data given below :

Class intrval :	0 – 10	10 – 20	20 – 30	30 – 40
Frequency :	1	3	4	2

Solution :

Class Interval	Freq f	Mid value x	$d = \frac{x-25}{10}$	fd	fd^2	fd^3	fd^4
0 - 10	1	5	–2	–2	4	– 8	16
10 - 20	3	15	–1	–3	3	–3	3
20 - 30	4	25	0	0	0	0	0
30 - 40	2	35	1	2	2	2	2
Total	10			–3	9	–9	21

Thus, $N = 10,\ \sum_{i=1}^{n} fd = -3\ ,\ \sum_{i=1}^{n} fd^2 = 9\ ,\ \sum_{i=1}^{n} fd^3 = -9$, and $\sum_{i=1}^{n} fd^4 = 21$

The raw moments of x about the arbitrary point A = 25 are given by

$$\mu_1' = h\frac{\sum_{i=1}^{n} fd}{N} = 10\left(\frac{-3}{100}\right) = -3$$

$$\mu_2' = h^2\frac{\sum_{i=1}^{n} fd^2}{N} = 100\left(\frac{9}{10}\right) = 90$$

The central moments are :

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 90 - (-3)^2 = 90 - 9 = 81$$

$$\mu_3 = \mu_3' - 3\mu_2' \mu_1' + 2(\mu_1')^3 = -900 - 3(90)(-3) + 2(-3)^3$$

$$= -900 + 810 - 54 = -144$$

$$\mu_4 = \mu_4' - 4\mu_3' \mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

$$= 21{,}000 - 4(-900)(-3) + 6(90)(-3)^2 - 3(-3)^4$$

$$= 21{,}000 - 108{,}00 + 4860 - 243 = 14817$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{14817}{(81)^2} = 2.258 \qquad \text{(Approx)}$$

Excess of kurtosis, $\gamma_2 = \beta_2 - 3 = 2.258 - 3 = -0.742$

This indicates that the given frequency distribution is platykurtic.

EXERCISE 5

1. Explain the term dispersion giving examples. Describe the need for measures of dispersion.
2. Explain different measures of dispersion, stating their merits and demerits.
3. Distinguish between absolute and relative measures of dispersion.
4. Define the term standard deviation. Explain its superiority over other measures of dispersion.
5. Explain the term standard deviation and root mean square deviation. Show that standard deviation is the least possible value of root mean square deviations.

10. Define the raw moments and central moments of a frequency distribution. Obtain the relationship between the r th central moment in terms of the raw moments.

11. What do you understand by skewness ? Explain various types of skewness and write their measures.

12. Explain the methods of measuring skewness and kurtosis in frequency distributions.

13. Show that for any distribution,

(i) the measure of kurtosis β_2 is greater than unity i.e. $\beta_2 \geq 1$

(ii) Bowley's coefficient of skewness is numerically less than unity.

14. Write a note on kurtosis.

15. The first four moments about the origin of a variable are respectively 1, 4, 10 and 46. Compute the first four central moments and find the coefficient of skewness and excess of kurtosis.

16. The first four moments of a distribution about the value 5 of a variable are respectively -4, 22, -117 and 560. Determine the moments about the mean.

17. For a frequency distribution, the mean is 10, the variance is 16, γ_1 is 1 and β_2 is 4. Obtain the first four moments about the origin.

18. Calculate the quatile deviation for the following data.

Class Interval :	5 -10	10 - 15	15 - 20	20 - 25	25 - 30
Frequency :	5	12	20	10	3

19. Find the standard deviation for the following data

Age in year :	21 - 25	26 - 30	31 - 35	36 - 40	41 - 45
No of Person :	15	20	28	17	12

20. Calculate the mean deviation about the mean and the coefficient of the mean deviation for the following data

Age :	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69
No of Person :	8	30	45	12	5

Class Interval :	0 - 2	2 - 4	4 - 6	6 - 8	8 - 10
Frequency :	4	16	13	7	5

23. The mean of two samples of sizes 40 and 60 are 25.7 and 35.2 and the standard deviations are 8 and 7 respectively. Find the mean and the standard deviation of the combined sample of size 100.

24. Calculate Karl Pearson's coefficient of skewness from the following series.

Wts.is kg. :	30-40	40-50	50-60	60-70	70-80	80-90	90-100,	100 and above
No. of person :	10	16	18	25	20	4	4	3

25. Find the mean and variance of the first 'n' natural numbers.

26. The sum of 20 observations is 300, the sum of the squares of these observations is 5000 and the median is 15. Find the coefficient of variation and the coefficient of skewness.

Fill in the blanks.

27. (i) The sum of the absolute deviations is minimum when measured from ______

(ii) Standard deviation is always ______ than the mean deviation from the mean.

(iii) The S.D of 100 observations is 12. If each observation is increased by 2, the SD of the new observations would be ______.

(iv) If $Q_1 = 10$ and $Q_3 = 30$ then the coefficient of quartile deviation is ______.

(v) For a symmetrical distribution, mean, median and mode ______.

(vi) For a distribution, if mean < mode, then the distribution is ______.

(vii) If for a frequency distribution, $\gamma_1 > 0$ and $\gamma_2 < 0$ then the distribution is ______ and ______.

(viii) If for a frequency distribution $\beta_2 > 3$, then its frequency curve is called ______.

(iv) Variance is always non-negative.

(v) Relative measures of dispersion are independent of units of measurement.

(vi) If the mean and the SD of a distribution are 20 and 4 respectively, then the C.V = 15%

(vii) For a distribution of 5 observations having values 4 each, the SD is 4.

ANSWERS

EXERCISE - 5

15. $\mu_2=3$, $\mu_3=0$, $\mu_4=27$, $\beta_1=0$, $\beta_2=3$

16. $\mu_1=0$, $\mu_2=6$, $\mu_3=19$, $\mu_4=32$

17. $\mu_1'=1$, $\mu_2'=116$, $\mu_3'=1544$, $\mu_4'=23184$

18. 3.5625 19. 4.17 20. M.D = 6.431

Coefficient of mean deviation about mean = 1.527 (Approx)

21. M.D = 21.38

Coefficient of M.D = 1.0975

22. 24.209 23. Mean 31.4, $\sigma = 8.755$

24. Sk = – 0.2155

25. Mean = $\frac{n+1}{2}$, variance $\frac{n^2-1}{12}$

26. C.V = 33.3, Sk = 0

27. (i) median (ii) greater (iii) 12 (iv) 0.5 (v) coincide (vi) negatively skewed (vii) positively skewed and platy kurtic (viii) leptokurtic (ix) k = $\frac{1}{3}$

28. (i) false (F) (ii) false (F) (iii) true (T) (iv) true (T) (v) true (T) (vi) false (F) (vii) false (F)

(iii) Range is based on all the observations of the data.

(iv) Coefficient of range is a measure of dispersion.

(b) Which of the following is correct ?

Coefficient of variation for a data with mean $\overline{x}$ and standard deviation σ is

(i) $\frac{\overline{X}}{\sigma}\times100$ (ii) $\frac{\overline{X}}{\sigma}\times1000$ (iii) $\frac{\sigma}{\overline{X}}\times100$ (iv) $\frac{\sigma}{\overline{X}}\times1000$

(c) Which of the following measures of variability is least affected by extreme items ?

(i) Range (ii) Mean Deviation (iii) Standard Deviation

(iv) Quartile Deviation

(d) Which of the following is the formula for quartile deviation ?

(i) $\frac{Q_1+Q_2}{2}$ (ii) $\frac{Q_1+Q_3}{2}$ (iii) $\frac{Q_3-Q_2}{2}$ (iv) $\frac{Q_3-Q_1}{2}$

(e) Which of the following is not true ?

(i) Excess of kurtosis, $\gamma_2 = \beta_2 - 3$.

(ii) Coefficient of skewness, $\gamma_2 = \mu_3 / \sigma^3$.

(iii) For a moderately skewed distribution, $S_k = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$.

(iv) Bowley's coefficient of skewness is equal to $(Q_3 + 2Q_2 - Q_1) / (Q_3 + Q_1)$.

of equal magnitude is zero.

(iv) The standard deviation of the first N natural numbers is $\sqrt{\frac{N^2+1}{12}}$

(g) Indicate which of the following is true.

(i) Interquartile range is the best measure of dispersion.

(ii) Variance and coefficient of variation have the same meaning.

(iii) Lorenz curve is a graphical representation of measure of skewness.

(iv) Standard deviation is the positive square root of the variance.

(h) State which of the following is not true.

(i) One of the purposes of studying dispersion is to determine reliability of an average.

(ii) One of the desirable requirments of a good measure of dispersion is that it should be based on all the items of the distribution.

(iii) Range is the simplest method of measuring dispersion.

(iv) Coefficient of range is calculated by using the formaula $\frac{L+S}{L-S}$ where L is the largest and S is the smallest item of the distribution.

(i) Which of the following is true ?

(i) Quartile deviation includes all the observations of the data.

(ii) Quartile deviation is amenable for further algebraic treatments.

(iii) Quartile deviation covers the central 80 % of the observations.

(iv) Quartile deviation is computed by using the central 50 % of the observations.

2. **(a) Fill in the blanks :**

(i) The sum of squares of deviations from ________ is minimum.

(ii) S.D is always ________ than the range.

(iii) The relation between standard deviation and mean deviation about mean is ________.

(iv) In a factory if more employees get less wage and less employee get more wage than the coefficient of skewness would re ________.

(v) The arithmetic mean and the standard deviation of the marks of a group of 100 students in a class are respectively 82 and 16. If 5 is subtracted from each of the marks, the values of the arithmetic mean and the standard deviations would be ________ and ________ respectively.

(b) Indicate True (T) or False (F) in each of the following questions :

(i) A more peak topped curve is called positively skewed curve.

(ii) Pearson's coefficient of skewness lies between -1 and $+1$.

(iii) For a frequency distribution the first central moment μ_1 is always equal to zero.

(iv) Relative measures of dispersion are always independent of units of measurement.

(v) To compare the consistency of two distributions their variances are compared.

3. **Give short answers to the following questions :**

(a) Indicate three requirments of a good measure of dispersion.

(b) State two algebraic properties of standard deviation.

(g) Write the demerits of standard deviation.

(h) State three tests of skewness.

(i) What is meant by kurtosis ?

(j) What is Lorenz Curve and for what purpose is it used ?

ANSWERS

1. (a) (iv)   (b) (iii)   (c) (iv)   (d) (iv)   (e) (iv)

   (f) (iii)   (g) (iv)   (h) (i)   (i) (iv)   (j) (iv)

2. (a) (i) Median   (ii) less   (iii) MD < SD   (iv) Negative   (v) 77, 16

   (b) (i) F   (ii) F   (iii) T   (iv) T   (v) F

★★★

6.1 INTRODUCTION

The need for statistical information seems endless in the present-day society. Particularly in developing countries, data are regularly collected to satisfy the need for information about various aspects like national income accounts, input-output tables, various production indices, price indices and a host of other quantitative indicators. It is very clear that without the relevant data, it is not possible to formulate policy objectives for a complex economy like ours. It is a fact that modern society is increasingly becoming an information oriented society. In this society, various economic and social processes are respresented by certain quantitative characteristics that require different kinds of information in the form of data.

The task of collecting data is becoming increasingly complex and difficult day-by-day. The total number of units to be consulted and investigated for obtaining the required information may be too large, while the resources in terms of money, time and manpower etc. may be limited. Moreover, obtaining error-free information from such a large scale investigation makes the job even more tedious. As a result, very often we try to obtain the required information from a smaller group that is easier to handle and control. However, it is important here to ensure that this smaller group which gives the required information, is truly representative of the entire collection of relevant units. The subject matter of sampling provides a mathematical theory for obtaining such kind of representative groups.

Some common examples where sampling is used are cited below :

(i) If a person wants to purchase a basket of oranges, he examines one or two from the basket and on the basis of the information gained from the ones examined, he makes his decision about the whole basket, i.e., decides whether to purchase or not.

6.2. CENSUS AND SAMPLE SURVEY:

In this section, we will distinguish between the census and the sampling methods of collecting data. We will also explain the meaning and coverage of census survey and sample survey.

6.2.1. Population and Census:

We have a collection of units defined according to the aims and objects of statistical enquiry. By a unit we mean an entity on which we can make observations according to a well-defined procedure. A unit may refer to a single individual or a group of individuals of the population. For example, we may define each student of a college as a unit or each family of a city as a unit or each packet of a manufactured item as a unit. The entire collection of such units is called a population or universe. For example, we may have a population of human beings, cattle, output of a particular manufactured product, industrial units, agricultural farms etc, Thus, a population may consist of units which are either animate or inanimate.

A population is either finite or infinite. If in a population the number of units is finite, it is called a finite population and if the number of units is infinite, it is called an infinite population. All the students of a college constitute a finite population. Similarly, drawing 10 balls sucessively without replacement from an urn containing 100 balls, is regarded as sampling from a finite population. On the other hand the population of fish in Chilika Lake, the population of temperatures of a place at different times or the population of the outcomes of independent trials of tossing a coin are examples of infinite populations. In fact, by infinite we mean indefinitely large. However, in practice, we will be concerned with a finite population.

on each and every member of the population. However, in order to obtain information about certain characteristic of the population, we need not always take recourse to a census because of some constraints like time, cost, man power etc. . In practice, we get quite satisfactory results by studying appropriate sample taken from the population. The procedure of obtaining a sample is known as sample survey. Thus, in a sample survey, we consider a representative part of the population as sample and use the same to infer about the entire population. It may be noted here that drawing conclusion for population on the basis of sample is regarded as an inductive process.

6.3. SOME BASIC CONCEPTS:

We explain below, some of the basic concepts necessary in sampling theory:

6.3.1 Parameter :

In a statistical inquiry, we are concerned with the identification of a population which can be done by studying its various characteristics. As such our interest lies in studying one or more characteristics of the population. A measure of such a characteristic is called a parameter. For example, we may be interested in the mean income of the people living in a certain region during a particular year. We may also like to know the standard deviation of the income of these people. Here, both mean and standard deviation are parameters.

Mathematically speaking, the value of a parameter is computed from all the observations constituting the population. If θ is a parameter that we want to obtain from the population consisting of units whose measurements are denoted by $Y_1, Y_2, \ldots, Y_N$, then $\theta = f\left(Y_1, Y_2, \ldots, Y_N\right)$, where N is the total number of units and is called size of the population. Among the parameters, most important and frequently used ones are the population mean $\overline{Y}$, the population total Y and the population variance σ^2. These parameters are defined by computed. Under the circumstances, we try to estimate the parameters on the basis of the information obtained from a sample drawn from the population. This information based on the sample is called a statistic. For example, sample mean, sample median and sample standard deviation are all statistics. Thus, it is clear that a statistic is calculated from the values of the units that are included in the sample and can be defined as a function of the sample values. Let $y_1, y_2, \ldots, y_n$ be a sample of n observations taken from a population with parameter θ. If t is a statistic that we want to compute from the sample values $y_1, y_2, \ldots, y_n$ taken from the population consisting of units $Y_1, Y_2, \ldots Y_N$ then $t = f(y_1, y_2, \ldots, y_n)$, where n is the size of the sample.

6.3.3. Estimator and Estimate

The basic purpose of a statistic is to estimate a population parameter. When a statistic is used to estimate a parameter, it is called an estimate. An esimator is a random variable that may assume different values from sample to sample taken from the same population. The values that the estimator takes are called estimates. If we use the formula of the sample mean given by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

for calculating a statistic to estimate the parameter $\bar{Y}$ (the population mean), then this formula will serve as an estimator. Using this formula if we get $\bar{y} = 20$, say, then this 20 is an estimate.

6.4. ADVANTAGES OF SAMPLE SURVEY OVER CENSUS SURVEY

In many situations, we undertake a sample survey in preference to a census survey because of the following important advantages:

(ii) Cost of data collection and supervision,

(iii) Cost of processing and analysing data, and

(iv) Cost of publication of results of the survey.

In this cost break-up, (ii) and (iii) are in the nature of variable costs, while (i) and (iv) are the fixed cost items. In a sample survey, items (ii) and (iii) will definitely be much smaller than for a census survey. Even though the cost of data collection per unit covered may be higher in a sample survey, the total cost of all units covered would be much smaller than that of a complete enumeration. Thus, it is usually imperative to resort to sampling rather than complete enumeration.

2. Reduced Time: There is considerable saving in time and labour since data are collected and summarised much faster in a sample survey than in a census survey. The sampling results can be obtained more rapidly and the data can be analysed much faster since a fraction of the whole data is collected and processed. This is an important advantage, particularly when the information is urgently needed in a small time gap.

3. Greater Accuracy: In a sample survey, investigation is carried out on a few units of the population. So more efficient, sincere and highly qualified investigators can be employed and given intensive training and careful supervision of the field work and processing of results is possible as the volume of work is reduced considerably in a sample survey. This would produce more accurate results than in complete census. Better job performance could be expected and ensured. Errors of certain types could be minimised in a sample survey.

4. Greater Scope : In general a sample survey has a greater and wider scope than a complete enumeration. The complete enumeration is impracticable, rather inconceivable if the survey requires highly trained personnel and more sophisticated and costly and /or hypothetical, like the population of the out comes of all the throws that may be made with a coin, sampling is the only course available. Again if the population consists of units which are destroyed in course of inspection, a complete enumeration does not help, e.g., when we want to know the average life in hours of a particular brand of electric bulb, complete enumeration is impossible because by the time our investigation is over, all the electric bulbs of the brand must be fused. In such situations sampling is the only alternative course available.

Despite the above advantages, sample surveys are not always preferred to census surveys, e.g., when time and money are not important factors for consideration or when detailed information is wanted for all the subclasses into which the population may be divided (i.e., about all the units of the population) or when the population size is not large, a complete enumeration may be more appropriate than any sampling procedure.

6.5. LIMITATIONS OF SAMPLING

Sampling theory has its own limitations and the advantages of sampling over complete enumeration can be derived only if

(i) the units of the population are drawn in a scientific manner,

(ii) an appropriate sampling technique is used, and

(iii) the sample adequately represents the population.

Besides, the other limitations of sampling are :

(i) Unless the sample survey is properly planned and carefully executed, the results obtained may not be reliable and may be misleading.

(ii) In the absence of the services of qualified, skilled and experienced personnel, adequate supervision, sophisticated equipments and statistical techniques for

The primary aim of sampling theory is to make sampling more effective so that the answer to a particular question is given in a valid, efficient and economical way. The theory of sampling is based on the following three basic principles which are helpful in fulfilling the aim.

1. Principle of Statistical Regularity

The principle of statistical regularity has its origin in the mathematical theory of probability. The principle can be explained as

"A moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group."

This principle stresses on the desirability and importance of selecting the sample at random and in large number.

2. Principle of Validity

This principle states that the sampling design should provide valid tests and estimates about the population parameters. The samples obtained by the method of probability sampling satisfy this principle.

3. Principle of Optimisation

This principle takes into account the desirablity of obtaining a sampling design which gives optimum results in terms of efficiency and cost. The reciprocal of sampling variance of an estimate provides a measure of its efficiency, while a measure of the cost of the design is provided by the total expenses incurred in terms of money and man hour. The principle of optimisation consists in

(i) achieving a given level of efficiency at minimum cost, and

(ii) obtaining maximum possible efficiency with given level of cost.

The planning stage consists of the following steps:

1. Objectives of the Survey : It is most important to define in clear and concrete terms the objectives of the survey. Unless this is done to the satisfaction of the user of the data and the statistician in charge of the survey, full utility of the survey results may not be achieved. Moreover, the user should ensure that these objectives are commensurate with available resources in terms of money, manpower and time limit required for the survey.

2. Defining the Population to be Sampled: The population, i.e., the aggregate of objects from which the sample is to be drawn, should be defined in clear and unambiguous terms. For example, to estimate the average yield per plot for a crop, it is necessary to define the size of the plot in clear terms. The sampled population (population to be sampled) should coincide with the target population (population about which information is required). The demographic, geographical, administrative and other boundaries of the population must be specified so that there should be no ambiguity regarding the coverage of the survey.

3. Determination of Sampling Frame and Sampling Units: For the purpose of selecting a sample it is necessary that the population to be sampled is subdivided into a finite number of distinct and identifiable units called sampling units. The sampling units must cover the entire population and those must be distinct, non-overlapping, i.e., each element of the population must belong to only one sampling unit. For example, in socio-economic survey for selecting people in a locality, the sampling unit might be an individual person, a family or a household.

It is also essential for the purpose of sampling to have a list of all sampling units of the target population. Such a list is called a 'frame' and provides the basis for the selection and identification of the units in the sample. Sometimes instead of a list of sampling units, maps and other acceptable materials are used as frame for identification of the sampling

4. Determination of the Data to be collected: The data should be collected in accordance with the objectives and the scope of the survey. Unnecessary and irrelevant data which would never be used subsequently should not be collected and no important or essential information should be left. Once the type or nature of data to be collected is decided upon, the next step would be to prepare the questionnaire or schedule through which the data are to be collected.

A questionnaire or a schedule is a set of questions designed to elicit information on a subject or a sequence of subjects. A questionnaire is filled in by the respondent himself, whereas the answers to various questions in a schedule are recorded by the investigator or enumerator on the basis of information gathered by interviewing the respondent. Generally, a draft questionnaire is first prepared and tried over a small group of individuals to discover any ambiguity or defect in framing the questions. This is called "pretest" or "pilot survey". If necessary, the questionnaire is revised and finalised in the light of the pretest. The questions should be brief, easily intelligible, unambiguous, practical and as far as possible objective type, and must not leave much scope for guessing on the part of the interviewer.

5. Determination of Method of Data Collection:

There are two methods, commonly employed, for collecting data from human populations. They are discussed below :

(i) Interview Method - In this method, the investigator goes from house-to-house and interviews the individuals personally. He asks the questions one by one and fills in the schedule on the basis of the information supplied by them.

(ii) Mailed Questionnaire Method - In this method, the questionnaire is mailed to the individuals who are required to fill in and return.

or instrument to be used and similar other things.

6. Selection of an Appropriate Sampling Design:

This is the most important step in planning a sample survey. Selection of an appropriate sampling design is made from among several alternatives. While making a decision about the particular sampling design to be adopted for survey, the investigator

(i) decides whether unrestricted random sampling or a variant of that is to be used in the survey under consideration;

(ii) chooses the flexible variables in the sample, if any, in an optimum manner;

and (iii) if necessary, decides upon the details of a pilot or exploratory survey for the main design.

An appropriate sampling design should take into account the objective of the survey, the type of sampled population, the cost implication, the time limit and the degree of accuracy required. Besides, any relevant practical considerations should also be taken into account.

7. Organisation of Field Work : The achievement of the objectives of a sample survey depends, to a large extent, on reliable field work. If field work is done honestly; sincerely and according to the instructions laid down and if there is careful supervision of the field staff, there remains no doubt about achieving the objectives of the survey. It is therefore necessary to make provisions for adequate supervisory staff for inspection of the field work.

The execution stage involves actual field work such as identification of the sampled population and the sampling units and collection of information from the sampling units through questionnaires or schedules.

8. Analysis and Reporting :

The analysis and reporting stage again involves the following steps:

(i) Scrutiny of Data : The filled-in questionnaires or schedules should carefully be scrutinised to find out whether the data furnished are plausible and whether data on different should be handled carefully for valid conclusions.

(iii) Tabulation of Data : Data collected are to be arranged in tabular form. For small scale survey manual tabulation may be possible, while for large scale survey machine tabulation is expected to be more economical and quicker. Now-a-days, computers are available for tabulation of data. As such, while drafting a questionnaire, the format should be so prepared that it can be processed through a computer.

(iv) Statistical Analysis : On the basis of data available, necessary estimates of the population values are obtained. Relevant characteristic measures are calculated from the collected data. Testing of hypotheses in certain cases are also carried out as a part of statistical analysis.

(v) Reporting and Conclusions : This is the final stage of the survey. A report incorporating detailed statement regarding all the stages of the survey should be prepared. The report should present all the collected statistical information in a neat tabular form. It should contain proper interpretation of data and the derived conclusion. Also recommendations, if any, made for the survey should be incorporated. It is a good parctice to report the technical aspects of the design of the survey in the presentation of the results, e.g, the types of estimators used and their margins of errors expected.

(vi) Storing of Information for future Surveys : At the completion of the survey, arrangements should be made for proper storing of the information so that it serves as a guide to the organisers for future surveys.

6.8. SAMPLING AND NON - SAMPLING ERRORS

The term error refers to the difference between the true value and the observed or approximated value. In any survey, errors are inevitable. These may occur at different stages of the survey due to a number of factors like (i) approximations in measurement (ii)

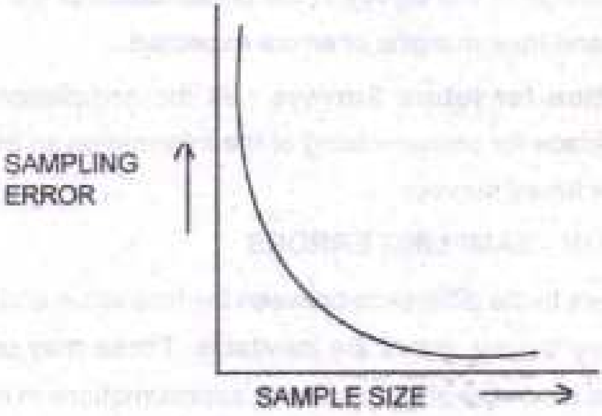
6.8.1. Sampling Earror:

The errors those arise due to the fact that only a sample is used to estimate the population parameters are termed as sampling arrors. Whatever may be the degree of cautiousness in selecting a sample, there will always be a difference between the population value (parameter) and its estimate. If θ be the parameter and t be an estmator of θ, then the absolute difference between these two, i.e., $|t - \theta|$ is called sampling error.

Sampling error is neither due to any lapse on the part of the respondent nor due to the investigator nor because of some such reasons. It arises because of the very nature of the procedure. It can never be completely eliminated. However, It can be reduced by increasing the size of the sample. In fact, the sampling error decreases with increase in the sample size. Sampling error is inversely proportional to the square root of the sample size. The relationship between sampling error and sample size can be illustrated graphically as is shown in Fig. 6.1. When the sample survey becomes a census survey, the sampling error becomes zero.

Fig. 6.1

Relationship between sample size and sampling error



2. Substitution : Investigators often substitute one convenient member of the population when difficulties arise in enumerating the appropriate sampling unit. This, obviously, leads to some bias since the characteristics possessed by the substituted unit are usually different from those possessed by the unit originally destined to be included in the sample.

3. Faulty demarcation of sampling units : Bias arises due to defective demarcation of sampling units in the area of the survey. These biases are common in the field of agricultural experiments or in crop cutting surveys. In such surveys, while dealing with border line cases, the investigator sometimes applies his own discretion whether to include or not to include a particular unit in the sample. Such decision influences the results either over estimating or under estimating the parameter.

4. Constant error due to wrong choice of the statistic : If a proper statistic is not taken to estimate the parameter, errors accumulate. For example, in estimating the population mean, the proper estimate is the sample mean. But instead of the sample mean if one uses the sample median, constant error will accumulate. Similarly, for estimating the population variance, the proper unbiased estimator would be the sample variance given by $S^2 = \Sigma(y_i - \bar{y})^2 / (n-1)$. But, if instead of S^2, the biased estimator of σ^2 given by $s^2 = \Sigma(y_i - \bar{y})^2 / n$ is taken, constant error is bound to arise.

8.8.2. Non-Sampling Error:

Besides sampling error, the sample estimate may be subject to other errors which, grouped togather, are termed non-sampling errors. Non-Sampling error, as the name suggests has nothing to do with sampling process. It primarily arises at different stages of the survey viz., observation, ascertainment and processing of the data and is thus present in both complete enumeration survey and sample survey. Various sources of non-sampling error are given below:

a survey of household expenditure, while some households maintain a meticulous diary on all expenses, the expenses become a matter of recall for others, resulting in measurement error. Measurement can be attributed to the following sources :

(i) The interviewer

(ii) The respondent

(iii) The questionnaire

(iv) The mode of interview, i.e. whether telephone, personal interview, self-administered questionnaire etc.

2. Error due to non-response : Non-response biases occur if full information is not obtained on all the sampling units. This happens mostly when questionnaires are sent to the respondents, but some of the respondents return the questionaires with incomplete answers or do not return them at all. This kind of attitude may be attributed to :

(a) the respondents are too casual to fill up the answers to the questions asked, or

(b) they are not in a position to understand the questions, or

(c) they do not like to disclose the information that has been sought.

It may be noted that the error due to non-response may also arise when data are collected through personal interviews. Here, this error may arise because some of the respondents

(a) may not like to give the information, or

(b) may not be available even after repeated visits.

3. Error due to inherent bias of the investigator : Every individual suffers from personal prejudices and biases. Despite the provision of the best possible training to the investigators, their personal biases may influence the interpretation of the questions asked significantly large because of the involvement of a large number of individuals in the data collection process. Sometimes, the non-sampling error becomes so large that it exceeds both the sampling and non-sampling errors taken together in a sample servey. Non-sampling errors can be minimised through

(a) a careful planning of the survey,

(b) providing proper training to the investigators, and

(c) making the questionnaire simple.

6.9. TYPES OF SAMPLING:

The technique or method of selecting a sample from a population is of fundamental importance in the theory of sampling and usually depends on the nature of the data and type of inquiry. The sampling procedures which are commonly used may be broadly classified under the following heads:

(i) Probability Sampling,

(ii) Non- probability Sampling, and

(iii) Mixed Sampling

6.9.1. Probability Sampling:

In probability sampling, the sampling units are selected according to some laws of chance, i,e, each unit in the population has some definite pre-assigned probability of being selected in the sample. It is also called random sampling and is based on the well established principles of probability theory. Simple random sampling, Stratified random sampling, Systematic sampling are some of the variants of random sampling.

6.9.2. Non-probability Sampling:

Non-probability sampling is based on the judgement or discretion of the person making a choice. Thus, in non-probability sampling, certain units may be selected example, a technical institute has to send 5 students for some managerial training in a company during the summer vacation. Initially, it may shortlist about 15 or 20 students who are considered to be suitable for the training by applying its, own discretion. Then from these shortlisted students, 5 students may finally be selected by employing random sampling procedure. Thus, the students selected will constitute a mixed sample.

6.10. SIMPLE RANDOM SAMPLING

The simplest and the most commonly used probability sampling is simple random sampling. In this sampling, each unit of the population has equal and independent probability of being selected in the sample.

There are two different schemes of simple random sampling, viz., simple random sampling with replacement (SRSWR) and simple random sampling without replacement (SRSWOR). The difference between these two schemes pertains to the way in which the sample units are selected. According to the procedure of simple random sampling with replacement, first a unit of the population is drawn at random, its features are noted and then the unit is replaced back to the whole lot so that the unit again becomes eligible for selection subsequently. Thus, the total number of units in the population always remains the same. In other words, the composition of the population remains unchanged, and each member of the population has the same chance or probability of being selected in the sample. In fact, if N is the size of the population, then the probability of selecting each unit of the population every time is 1/N. On the other hand, in case of simple random sampling without replacement, the unit once selected is not returned to the population in the sense that it becomes ineligible for selection again. As a result, after each successive draw, the composition of the population changes, i.e., one unit is eliminated from the total number of units of the population. Therefore for subsequent draw from the population the probability of any particular unit being picked up also gets changed. Suppose, the population size is N and we want to draw a sample of size n from it by the principle of SRSWOR.

It should be noted here that, from a population of size N, the number of samples of size n that can be drawn with replacement is N^n and so the probability of selecting a sample is $1/N^n$, while the number of samples that can be drawn without replacement is $\binom{N}{n}$ and probability of selecting a sample is $1/\binom{N}{n}$.

Example 6.1

Supose a population consists of the following 4 units (2,5,6,8). How many samples of size 2 can be drawn from it ?

(i) If we follow the procedure of SRSWR, the number of samples that can be selected is $N^n=4^2 = 16$.

The possible samples are as follows:

(2, 2), (2, 5), (2, 6), (2, 8), (5, 2), (5, 5) (5, 6), (5, 8),

(6, 2), (6, 5), (6, 6), (6, 8), (8, 2), (8, 5), (8, 6), (8, 8)

We should note that in sampling with replacement the order in which the units are selected also matters. Thus, (2, 5) and (5, 2) are considered as two different samples.

(ii) If we follow the procedure of SRSWOR, the number of samples that can be selected is $\binom{N}{n}=\binom{4}{2}=\frac{4!}{2!2!}=\frac{4\times3}{2}=6$

The possible samples are listed below :

(2, 5), (2, 6), (2, 8), (5, 6), (5, 8), (6, 8).

It may be noted that in sampling without replacement, once a member is selected it cannot be selected again for inclusion in the same sample. Thus, sample like (2, 2), (5, 5) etc., cannot be considered for selection. Similarly, if a sample like (2, 5) is selected, then another sample like (5, 2) cannot be selected.

(r -1) draws and (b) the probability that it is selected at the rth draw with the condition that it is not selected in the previous (r - 1) draws.

The probability under (a) is given by

(the probability that it is not selected at the first draw)x (the probability that it is not selected at the second draw)x.....x. (the probability that it is not selected at the (r - 1)th draw).

$$= \frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \frac{N-3}{N-2} \cdots\cdots \frac{N-r+1}{N-r+2} = \frac{N-r+1}{N}$$

The probability under (b) is given by 1/(N-r+1). Hence the required probability is (a)x(b) = 1/N, which is independent of r, the number of draws.

Again, the probability of selecting any unit of the population at the first draw is 1/N. Hence the theorem.

Theorem 6.2 The probability of a specified unit of a population of size N being included in a sample of size n is equal to n/N.

Proof : Since the specified unit can be included in the sample of size n either in the first or in the second or .... or in the nth draw, the event can happen in n mutually exclusive ways. So, the probability that the specified unit is included in the sample is the sum of the probabilities of these n mutually exclusive events. But, we know from Theorem 6.1 that the probability of selecting a specified unit of the population of size N at any draw is equal to 1/N. Thus, the required probability is

$$\frac{1}{N} + \frac{1}{N} + \ldots\ldots + \frac{1}{N}(\text{n times}) = \frac{n}{N}.$$

Hence the theorem.

[Proof of the above two theorems under SRSWR scheme are left as an exercise] possible sample of a given size has the same chance of selection. That is, each unit of the population is equally likely to be chosen at any stage of the sampling process and selection of one unit in no way influences the selection of another unit of the population.

There are two most commonly used methods available to draw a simple random sample. The first is 'Lottery method' and the second is 'Random Numbers Table method'. Irrespective of the method we decide to use, every unit of the sampling frame should be assigned with a separate identifying number.

6.11.1. Lottery Method : The simplest method of drawing a simple random sample is the lottery method, which is illustrated below by means of an example.

Suppose we want to select a simple random sample of size n out of a population of size N. We assign the numbers 1 to N; one number to each of the population units and write these numbers on N different slips. These slips must be homogeneous with respect to each other in shape, size and colour. These slips are then put in a container and thoroughly suffled. Finally, n slips are drawn out of the container one by one. The n units corresponding to the numbers appearing on the slips drawn constitute a simple random sample.

The disadvantages of the lottery method are :

(i) making homogeneous slips for writing the identification numbers of the units of the population is a tedious affair, especially when the population size is large and (ii) the quality of the sample depends on how thoroughly the slips have been mixed and how honestly they have been picked up.

6.11.2. Random Numbers Table Method :

The random numbers are a collection of numbers generated through a probability mechanism by using the digits from 0 to 9. The random numbers have the following properties:

with equal probability, we choose the units with the help of Random Numbers Tables. Such tables have been prepared by different persons like Kendall and Smith (1939), Tippet (1927), Fisher and Yates (1938) Rao, Mitra and Mathai (1966) etc.

Random number tables have been prepared from different sources like the British Census Reports, Thomson's 20-figure logarithmic tables etc. The randomness of the numbers in the table have been established through various statistical tests. Of the different types of random numbers tables, the table due to Tippet is very popular and is commonly used for various purposes. We give below in Table 6.1 an extract of the Tippet's table of random numbers as a model.

Table - 6.1

Extract from Tippet's Table of Random Numbers (Four digits).

2952	6641	3992	9792	7979	5911	3170
4167	9524	1545	1396	7203	5356	1300
2370	7583	3408	2762	3563	1089	6913
0560	5246	1112	6107	6008	8126	4233
2754	9143	1405	9025	7002	6111	8816

An extract of another random numbers table is also given in the Appendix - I.

The procedure of selection of a simple random sample by using RNT consists in the following steps :

1. Identify all the N units in the population with the number from 1 to N irrespective of the order of units in the population.

2. Specify the starting point and the direction of movement. For example, to select three

the required number of sample units is obtained. These selected random numbers refer to the serial numbers of the units of the population and are included in the sample.

The following example will illustrate the procedure :

Example 6.2 :

Draw a simple random sample (without replacement) of 15 schools from a list of 338 schools.

Since there are 338 schools in all, we allot a number from 001 through 338 to each of the schools for identification. Then we start from the first column on the extreme left of the random numbers table given in the Appendix - I and go on selecting 3-digit numbers moving downward, discarding those which are either greater than 338 or repeated. Thus, we have the numbers

125, 326, 012, 237, 035, 251, 165, 131, 198, 033, 161, 209, 051, 052 and 331

Therefore, the schools assigned with these serial numbers constitute a simple random sample of size 15 taken from the list of 338 schools.

The procedure rejects a large number of random numbers. A device commonly used to avoid the rejection of such large number of random numbers is to divide a 3-digit random number greater than 338 by 338 and to choose the serial numbers from 001 through 338 equal to the remainders. When the remainder is zero, it refers to the serial number 338. However, it is necessary to reject the random numbers from 677 to 999. In case the numbers from 677 to 999 are included, the schools bearing serial numbers from 001 to 323 would get larger chance of selection i.e, equal to 3/999, while those with serial numbers from 324 to 338 get a chance equal to 2/999. If we use this procedure and the same table given in Appendix I, the 3-digit random numbers would give the following serial numbers:

125, 206, 326, 193, 012, 237, 035, 251, 325, 338, 114, 231, 078, 112 and 126.

given by Y_i (i = 1, 2,......, N), we will have n values taken out of the N values. We will denote these n values by y_1, y_2, ..., y_n. It may be noted here that the capital letters represent the population values, while the small letters, the sample values. Further, the 'y' values are not different from those of the 'Y' values. The only difference is y_1, y_2,...., y_n do not exactly correspond to Y_1, Y_2,...., Y_n chronologically but are any randomly chosen n values out of the values Y_1, Y_2,...., Y_N

We consider the following notations :

Population mean, $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ ($\overline{Y}$ is also called the mean per unit of the population.)

The population total, $Y = \sum_{i=1}^{N} Y_i$

The population mean square, $S^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} Y_i^2 - N\overline{Y}^2\right]$

The population variance, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \frac{N-1}{N}S^2$

The samiple mean, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

The sample mean square, $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right]$

6.13 UNBIASEDNESS PROPERTY OF THE SAMPLE MEAN, THE SAMPLE MEAN SQUARE AND THE VARIANCE OF THE SAMPLE MEAN.

Theorem 6.1: In simple random sampling without replacement (SRSWOR), the sample mean is an unbiased estimator of the population mean, i.e., E $(\overline{y}) = \overline{Y}$.

i.e., $\sum_{i=1}^{n} y_i = \sum_{i=1}^{N} a_i Y_i$,

where $a_i = \begin{cases} 1, & \text{if the ith unit of the population is included in the sample} \\ 0, & \text{if the ith unit of the population is not included in the sample} \end{cases}$

So, $$\bar{y} = \frac{1}{n}\sum_{i=1}^{N} a_i Y_i \qquad \text{..... (6.1)}$$

Taking expectation on both the sides, we get

$$E(\bar{y}) = E\left[\frac{1}{n}\sum_{i=1}^{N} a_i Y_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{N} E(a_i) Y_i \qquad \text{..... (6.2)}$$

Now, $E(a_i) = 1.\ P(a_i = 1) + 0.P\ (a_i = 0)$

= 1. P (that the ith unit is included in the sample)

+ 0.P (that the unit is not included in the sample)

$$= 1.\ \frac{n}{N} + 0\ (1 - \frac{n}{N})$$

$$= \frac{n}{N} \qquad \text{.....(6.3)}$$

Hence, substituting (6.3) in (6.2), we find

$$E\ (\bar{y}) = \frac{1}{n}\sum_{i=1}^{N} \frac{n}{N} Y_i$$

Theorem 6.2 In simple random sampling without replacement (SRSWOR), the sample mean square is an unbiased estimator of the population mean square, i.e.,

$$E\,(s^2) = S^2$$

Proof. We have

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(y_i^{\,2} - n\bar{y}^2)$$

$$\therefore \qquad = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}y_i^{\,2} - \bar{y}^2\right) \qquad \text{..... (6.5)}$$

Now, $$\bar{y}^2 = (\frac{1}{n}\sum_{i=1}^{n}y_i)^2 = \frac{1}{n^2}(\sum_{i=1}^{n}y_i^{\,2} + \sum_{i\neq j=1}^{n}y_i y_j)$$

$$s^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}y_i^{\,2} - \frac{1}{n^2}\left\{\sum_{i=1}^{n}y_i^{\,2} + \sum_{i\neq j=1}^{n} y_i y_j\right\}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}y_i^{\,2} - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} y_i y_j \qquad \text{..... (6.6)}$$

Defining a_i as in Theorem 6.1, we have

$$E(s^2) = \frac{1}{n}\sum_{i=1}^{N}E(a_i)Y_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{N}E(a_i a_j)Y_i Y_j \qquad \text{......(6.7)}$$

Now, $$E(a_i) = 1.P(a_i{=}1) + 0.P(a_i{=}0) = \frac{n}{N} \qquad \text{...... (6.8)}$$

$$= \frac{n}{N}.\frac{n-1}{N-1} \quad \text{(since samples are taken without replacement)} \qquad \text{...... (6.8a)}$$

Substituting (6.8) and (6.8a) in (6.7), we have

$$E(s^2) = \frac{1}{n}.\sum_{i=1}^{N}\frac{n}{N}Y_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{N}\frac{n(n-1)}{N(N-1)}.Y_iY_j$$

$$= \frac{1}{N}\sum_{i=1}^{N}Y_i^2 - \frac{1}{N(N-1)}\sum_{i\neq j=1}^{N}Y_iY_j$$

$$= S^2 \qquad \text{[by (6.6)]}$$

Hence the theorem.

Theorem 6.3 : In simple random sampling without replacement (SRSWOR), the variance of the sample mean is given by

$$\text{Var}\,(\overline{y}) = \frac{N-n}{N}.\frac{S^2}{n} = (1-f)\frac{S^2}{n}, \qquad \text{...... (6.9)}$$

where f = n/N is the sampling fraction. Furthermore, an unbiased estimator of $\text{Var}\,(\overline{y})$ is

$$\hat{V}\,(\overline{y}) = \frac{N-n}{N}.\frac{s^2}{n} = (1-f)\frac{s^2}{n}, \qquad \text{...... (6.10)}$$

Proof.

We have by definition

$$\text{Var}\,(\overline{y}) = E(\overline{y}^2) - [E(\overline{y})]^2$$

$$= E(\overline{y}^2) - (\overline{Y}^2) \qquad \text{...... (6.11)}$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n} E(y_i^2) + \sum_{i\neq j=1}^{n} E(y_i y_j)\right]$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} E(a_i)Y_i^2 + \sum_{i\neq j=1}^{N} E(a_i a_j)Y_i Y_j\right]$$

Substituting from (6.8) and (6.8a) in (6.12), we get

$$E(\bar{y}^2) = \frac{1}{n^2}\left[\frac{n}{N}\sum_{i=1}^{N} Y_i^2 + \frac{n(n-1)}{N(N-1)}\sum_{i\neq j=1}^{N} Y_i Y_j\right] \qquad ......(6.13)$$

But, $\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2$

$$\Rightarrow \qquad \sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 + N\bar{Y}^2 = (N-1)S^2 + N\bar{Y}^2 \qquad ......(6.14)$$

Again since, $\left(\sum_{i=1}^{N} Y_i\right)^2 = \sum_{i=1}^{N} Y_i^2 + \sum_{i\neq j=1}^{N} Y_i Y_j$

we have, $\sum_{i\neq j=1}^{N} Y_i Y_j = \left(\sum_{i=1}^{N} Y_i\right)^2 - \sum_{i=1}^{N} Y_i^2$

$$= (N\bar{Y})^2 - \sum_{i=1}^{N} Y_i^2$$

$$= N^2\bar{Y}^2 - (N-1)S^2 - N\bar{Y}^2 \quad \text{(using 6.14)}$$

$$= (N-1)\left[N\bar{Y}^2 - S^2\right] \qquad ......(6.15)$$

$$\mathrm{Var}(\bar{y}) = \frac{N-n}{N} \cdot \frac{S^2}{n}.$$

The proof for the second part of the theorem is straight forward, for it follows from the fact that the sample mean square s^2 is an unbised estimator of S^2.

Thus, $$E\left[\frac{N-n}{N}\frac{s^2}{n}\right] = \frac{N-n}{Nn}E(s^2) = \frac{N-n}{N} \cdot \frac{S^2}{n} = \mathrm{Var}(\bar{y}).$$

Corollary 6.1 : $\hat{Y} = N\bar{y}$ is an unbiased estimator of the population total Y and its variance is given by

$$\mathrm{Var}\ (N\bar{y}) = N^2\ \mathrm{Var}\ (\bar{y}) = N^2 \frac{N-n}{Nn} S^2 = \frac{N(N-n)}{n} S^2$$

Thus, $$\mathrm{Var}(\hat{Y}) = N(N-n)\frac{S^2}{n} = N^2(1-f)\frac{S^2}{n} \qquad \text{.....(6.17)}$$

If we let N tend to infinity in Theorem 6.3, we obtain the following result for simple random sampling with replacement.

Corollary 6.2 : In simple random sampling with replacement (SRSWR), i.e., sampling from an infinite population the sample mean $\bar{y}$ is an unbiased estimator of $\bar{Y}$ and its variance is given by

$$\mathrm{Var}\ (\bar{y}) = \frac{\sigma^2}{n} \qquad \text{..... (6.18)}$$

Further, an unbiased estimator of the variance is given by

$$\hat{V}(\bar{y}) = \frac{s^2}{n}. \qquad \text{.....(6.19)}$$

$$1-\frac{n-1}{N-1} \text{ or } \frac{N-n}{N-1}$$

When n is very small as compared to N, the finite population correction factor does not differ much from unity and the variance of the sample mean becomes approximately equal to the variance of the sample mean of a sample drawn from an infinite population.

(2) The standard error (SE) of the sampling distribution of $\bar{y}$ is given by

$$SE(\bar{y}) = \sqrt{\frac{N-n}{N}} \cdot \frac{S}{\sqrt{n}}.$$

When S^2 is not known, we replace S^2 by its unbiased estimate s^2. Thus, we get

$$\text{Est. } S\,E(\bar{y}) = \sqrt{\frac{N-n}{N}} \cdot \frac{s}{\sqrt{n}}. \qquad \text{...... (6.20)}$$

(3) For simple random sampling with replacement from a finite population,

$$\text{Var}\,(\bar{y}) = \frac{\sigma^2}{n} = \frac{N-1}{N} \cdot \frac{S^2}{n}. \qquad \text{..... (6.21)}$$

Comparing (6.9) with (6.21) we find that Var $(\bar{y})_{srswor}$ < Var $(\bar{y})_{srswr}$, i.e., SRSWOR provides a more efficient estimator of $\bar{Y}$ as compared to SRSWR.

Example 6.3

A population consists of six units with values 5, 6, 8, 9,10 and 12. List down all possible samples of size 3 taken without replacement and show that the sample mean is an unbised estimate of the population mean.

4	5, 6, 12	23	23/3
5	5, 8, 9	22	22/3
6	5, 8, 10	23	23/3
7	5, 8, 12	25	25/3
8	5, 9, 10	24	24/3
9	5, 9, 12	26	26/3
10	5, 10, 12	27	27/3
11	6, 8, 9	23	23/3
12	6, 8, 10	24	24/3
13	6, 8, 12	26	26/3
14	6, 9, 10	25	25/3
15	6, 9, 12	27	27/3
16	6, 10, 12	28	28/3
17	8, 9, 10	27	27/3
18	8, 9, 12	29	29/3
19	8, 10, 12	30	30/3
20	9, 10, 12	31	31/3
	Total	**500**	**500/3**

$$\therefore \qquad E(\bar{y}) = \frac{1}{\binom{N}{n}} \Sigma\, \bar{y} = \frac{1}{20} \times \frac{500}{3} = \frac{25}{3} = 8.3$$

Now, $$\overline{Y} = \frac{1}{N} \Sigma\, Y_i = \frac{1}{6}(5+6+8+9+10+12)$$

$$= \frac{50}{6} = 8.3$$

(i) defective items in a large consignment of such items,

(ii) children below 5 years vaccinated against polio in a certain place,

(iii) the educated unemployed persons in a city and so on.

In such cases, every sampling unit in the population is placed in one of the two classes C or C', respectively, according as it possesses or does not possess the given attribute.

Let there be N units in the population out of which 'A' units belong to class C and the remaining $N - A = A'$ units belong to class C'. Suppose in a simple random sample of n units taken out of the population, 'a' units are found to belong to class C. We define

$P = \frac{A}{N}$, the proportion of units in the population possessing the given attribute.

$Q = \frac{N-A}{N} = 1 - P$, the proportion of units in the population which donot possess the given attribute.

$p = \frac{a}{n}$, the proportion of units in the sample possessing the given attribute.

and $q = \frac{n-a}{n} = 1 - p$, the proportion of units in the sample which do not possess the given attribute.

With the ith sampling unit, let us associate a variable Y_i (i =1, 2......N) that assumes the values 1 or 0 according as the unit belongs to C or C', respectively. Similarly, let y_i (i=1, 2, ...n) be associated with the ith sampled unit and assumes values 1 or 0 according as the ith sampled unit possesses the given attribute or not.

N

and $$\bar{y} = \frac{a}{n} = p \qquad .....(6.23)$$

Similarly, we have

$$\sum_{i=1}^{N} Y_i^2 = A = NP$$

and $$\sum_{i=1}^{n} y_i^2 = a = np$$

$$\therefore \qquad S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right]$$

$$= \frac{1}{N-1}\left[NP - NP^2\right]$$

$$= \frac{NP(1-P)}{N-1} = \frac{NPQ}{N-1} \qquad ..... (6.24)$$

and $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{npq}{n-1} \qquad ..... (6.25)$$

Theorem 6.4 : In simple random sampling without replacement (SRSWOR), the sample proportion 'p' is an unbiased estimator of the population proportion P, i.e.,

$$E(p) = P$$

Proof. We know from (6.4) that, in SRSWOR the sample mean is an unbiased estimator of the population mean, i.e,

$$E(\bar{y}) = \bar{Y}$$

$$\text{Var}(p) = \frac{(N-n)}{N-1} \cdot \frac{PQ}{n} \qquad \text{......(6.27)}$$

Proof. We know that $\bar{y} = p$

Hence, $\text{Var}(p) = \text{Var}(\bar{y})$

$$= \frac{(N-n)}{Nn} \cdot S^2$$

$$= \frac{(N-n)}{Nn} \times \frac{NPQ}{N-1} \qquad \text{(using 6.24)}$$

$$= \frac{(N-n)}{N-1} \cdot \frac{PQ}{n} \qquad \text{(Proved)}$$

Corollary 6.4 : In SRSWOR, an unbiased estimator of Var (p) is given by

$$\hat{V}(p) = \frac{N-n}{N(n-1)} pq \qquad \text{......(6.28)}$$

Remarks : If N is sufficiently large as compared to n, i.e., when fpc. is ignored, then (6.28) reduces to

$$\hat{V}(p) = \frac{pq}{n-1}, \qquad \text{......(6.29)}$$

However, for large samples, $(n-1) \simeq n$ and we get

$$\hat{V}(p) \simeq \frac{pq}{n}, \qquad \text{......(6.30)}$$

a formula which is commonly used in practice.

and (ii) the confidence coefficient with which we want this estimate to lie within the permissible margin of error.

Let us consider that the parameter $\overline{Y}$, the population mean, is to be estimated. We know that $\overline{y}$, the sample mean of a sample of size n is an unbiased estimator of $\overline{Y}$. If the permissible error in estimating is 'd' and confidence coefficient is $1-\alpha$, then the sample size n is determined by the equation

$$P\left[|\overline{y}-\overline{Y}|<d\right]=1-\alpha \qquad \text{......(6.31)}$$

$$\text{or} \quad P\left[|\overline{y}-\overline{Y}|\geq d\right]=\alpha, \qquad \text{...... (6.32)}$$

where α is very small pre-assigned probability and is known as the level of significance.

If n is sufficiently large and we consider SRSWOR, then the statistic

$$Z=\frac{\overline{y}-E(\overline{y})}{SE(\overline{y})}=\frac{\overline{y}-\overline{Y}}{S\sqrt{\frac{1}{n}-\frac{1}{N}}} \qquad \text{.......(6.33)}$$

follows standard normal distribution. If Z_α is the α100% point of the standard normal distribution, then we have

$$P\left[\frac{|\overline{y}-\overline{Y}|}{S\sqrt{\frac{1}{n}-\frac{1}{N}}}\geq Z_\alpha\right]=\alpha \qquad \text{.....(6.34)}$$

$$\Rightarrow \qquad n = \frac{\left(\frac{Z_\alpha}{d} S\right)^2}{1 + \frac{1}{N}\left(\frac{Z_\alpha S}{d}\right)^2} \qquad \text{.... (6.36)}$$

which depends on the population parameter S. In actual practice, either a guessed value or an estimated value of S^2 undertaking a pilot survey or from past survery of similar population may be used.

Note : Although in most texts the symbol Z_α is used for the value of Z beyond which $(\alpha/2)$% of the right tail area lie under normal probability curve, it is better to replace Z_α by $Z_{\alpha/2}$ and $-Z_\alpha$ by $-Z_{\alpha/2}$ so that

$$P\,(|Z| \ge Z_{\alpha/2}) = \alpha .$$

Using this symbol, n can be determined. by the formula

$$n = \frac{(Z_{\alpha/2} S / d)^2}{1 + \frac{1}{N}\left(\frac{Z_{\alpha/2} S}{d}\right)^2}$$

In the event of n being small, instead of using Z, the statistic t is used. The application of t statistic is beyond the scope of the present text.

2. Using simple random sampling, we can select an unbiased sample. This is due to the fact that in SRS, each unit of the population is provided with equal opportunity for inclusion in the sample. As a result, the bias due to human preferences or judgement is completely eliminated.

3. Through estimation of sampling error, we can assess the accuracy of the results and specify the efficiency.

4. If the population size is not too large, simple random sampling is a simple, convenient and easily implementable sampling procedure.

Disadvantages :

1. The selection of a simple random sample requires an up-to-date frame, which is usually not made available. Moreover, preparation of an up-to-date frame is a difficult task. This restricts the use of simple random sampling technique.

2. A simple random sample may result in the selection of the sampling units which are scattered over a large geographical area. As a result, the cost of collecting the data may be much in terms of time and money.

3. For a given precision, simple random sampling usually requires larger sample size as compared to stratified random sampling which is discussed in section 6.17.

4. If homogeneity property of the units of the population is not attained, then simple random sampling gives rise to larger variability of the estimates resulting in reduced precision.

Example 6.4

Consider a population consisting of 5 units : 4, 7, 8, 1 and 10. Suppose a sample of 2 units is to be selected from it by the method of simple random sampling without replacement. We want to obtain the sampling distribution of the sample mean and show that the sample and size of the sample, n = 2

The population values are $Y_1 = 4$, $Y_2 = 7$, $Y_3 = 8$, $Y_4 = 1$ and $Y_5 = 10$

The number of samples that can be selected without replacement is

$$\binom{N}{n} = \binom{5}{2} = 10$$

All possible samples of size 2 taken without replacement along with the corresponding sample means ($\bar{y}$) are presented in the following table :

Table 6.3

Possible Samples and Sample Means

Sample	Sample Mean ($\bar{y}$)
(4, 7)	5.5
(4, 8)	6.0
(4, 1)	2.5
(4, 10)	7.0
(7, 8)	7.5
(7, 1)	4.0
(7, 10)	8.5
(8, 1)	4.5
(8, 10)	9.0
(1, 10)	5.5

The frequency distribution of the sample means is given in the following table :

4.5	1
5.5	2
6.0	1
7.0	1
7.5	1
8.5	1
9.0	1

From the above table, we can form the probability distribution of the sample mean (or sampling distribution of the sample means) as follows :

Table 6.5

Sampling Distribution of Sample Means

Sl. No	Sample mean ($\bar{y}$)	Probability (p)
1	2.5	1/10
2	4.0	1/10
3	4.5	1/10
4	5.5	2/10
5	6.0	1/10
6	7.0	1/10
7	7.5	1/10
8	8.5	1/10
9	9.0	1/10

Hence, E ($\bar{y}$) = Average of all sample means

$$\overline{Y} = \frac{4+7+8+1+10}{5} = \frac{30}{5} = 6$$

Thus, $E(\overline{y}) = \overline{Y}$. This shows that $\overline{y}$ is an unbiased estimator of $\overline{Y}$.

The variance of $\overline{y}$ is

$$\begin{aligned} \text{Var}\,(\overline{y}) &= E(\overline{y} - \overline{Y})^2 \\ &= E(\overline{y})^2 - \overline{Y}^2 \\ &= \sum_{i=1}^{9} \overline{y}_i^2 p_i - \overline{Y}^2 \\ &= 39.75 - 3 \\ &= 3.75 \end{aligned}$$

According to the formula

$$\begin{aligned} \text{Var}\,(\overline{y}) &= \frac{N-n}{N} \cdot \frac{S^2}{n} \\ &= \frac{5-2}{5 \times 2}\left[\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2\right] \\ &= \frac{3}{10 \times 4}\left[\sum_{i=1}^{5} Y_i^2 - 5\overline{Y}^2\right] \\ &= \frac{3}{40}\left[230 - 5 \times 36\right] \\ &= 3.75 \end{aligned}$$

population given in Example 3.1, show that

$$\text{(i)} \quad E(\bar{y}) = \bar{Y}$$

$$\text{and} \quad \text{(ii)} \quad Var(\bar{y}) = \frac{\sigma^2}{n}.$$

Solution :

In case of simple random sampling with replacement the number of possible samples is $5^2 = 25$, the probability of selection of each sample being $\frac{1}{25}$. All possible samples of size 2 along with sample means are listed in the table given below :

3	(4, 8)	12/2	144/4
4	(4, 1)	5/2	25/4
5	(4, 10)	14/2	196/4
6	(7, 4)	11/2	121/4
7	(7, 7)	14/2	196/4
8	(7, 8)	15/2	225/4
9	(7, 1)	8/2	64/4
10	(7, 10)	17/2	289/4
11	(8, 4)	12/2	144/4
12	(8, 7)	15/2	225/4
13	(8, 8)	16/2	256/4
14	(8, 1)	9/2	81/4
15	(8, 10)	18/2	324/4
16	(1, 4)	5/2	25/4
17	(1, 7)	8/2	64/4
18	(1, 8)	9/2	81/4
19	(1,1)	2/2	4/4
20	(1, 10)	11/2	121/4
21	(10, 4)	14/2	196/4
22	(10, 7)	17/2	289/4
23	(10, 8)	18/2	324/4
24	(10, 1)	11/2	121/4
25	(10, 10)	20/2	400/4
	Total	300/2	4100/4

$$= 41 - 36 = 5.$$

Further, $$\frac{\sigma^2}{n} = \frac{1}{n} \times \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

$$= \frac{1}{2}\left[\frac{1}{N}\sum Y_i^2 - \bar{Y}^2\right] = \frac{1}{2}\left[\frac{1}{5}\times 230 - 36\right] = \frac{10}{2} = 5$$

Thus, $$\text{Var}(\bar{y}) = \frac{\sigma^2}{n}.$$

Example 6.6

A random sample of 10 students was selected from a class consisting of 120 students following SRSWOR scheme. The marks secured by these 10 students in a test were as follows :

7, 4, 8, 10, 8, 6, 7, 9, 6, 5

Estimate the average marks of all the students of the class and the standard error of the estimate.

Solution :

Here, N = 120 and n = 10

The sample mean is computed as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{70}{10} = 7$$

Thus, the estimated average marks of all the students of the classs is 7.

$$Nn \quad n-1 \left[{}_{i=1} \qquad \right]$$

$$= \frac{120-10}{120\times 10}\cdot\frac{1}{9}\left[520-10\times 49\right]$$

$$= \frac{33}{108} = 0.31$$

Hence, $\qquad$ Est. SE. $(\bar{y}) = \sqrt{0.31} = 0.56.$

Example 6.7

A simple random sample (without replacement) of 15 households is drawn from a village consisting of 250 households. The number of persons in the households of the sample are given below:

5, 8, 6, 7, 5, 7, 8, 8, 9, 6, 4, 5, 3, 4, 5

Estimate the total number of people living in the village and obtain estimate of standard error.

Solution :

Here, N = 250, n = 15

The sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{90}{15} = 6$

Thus, $\qquad \hat{Y} = N.\bar{y} = 250\times 6 = 1250$

The total number of people in the village is estimated to be 1250.

Now, $\qquad$ Est. Var $(\hat{Y}) = \frac{N(N-n)}{n}s^2 = \frac{N(N-n)}{n}\times\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$

$$= \frac{N(N-n)}{n}\cdot\frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2 - n\bar{y}^2\right]$$

Example 6.8

From a list of 1200 students of a college, a simple random sample of 100 students showed that 45 were against shifting the date of examination. Estimate the total number of students against shifting the date of examination. Also find the standard error of your estimate.

Solution :

Here, $N = 1200$, $n = 100$ and $a = 45$.

$$\therefore \quad p = \frac{a}{n} = \frac{45}{100}$$

We know that $E(p) = P$

$E(Np) = NP = A$.

i.e., Np provides an unbiased estimate of A.

Thus, the estimated value of A, $\hat{A} = Np$

$$= 1200 \times \frac{45}{100} = 540$$

Hence, the total number of students who are against shifting the date of the examination is 540.

Further, $$\text{Var}(p) = \frac{N-n}{N-n} \cdot \frac{PQ}{n}.$$

An unbiased estimate of Var (p) is given by

$$\text{Est. Var}(p) = \frac{(N-n)}{N(n-1)}\, pq$$

Example 6.9

By complete enumeration of 500 units of a population, the mean and the variance were found to be 24 and 54.6, respectively. With simple random sampling, how large the sample must be taken so that the difference between the sample mean and the population mean would be less than 10% the population mean with 95% confidence?

Soultion :

Here, we are given : $N = 500$, $\overline{Y} = 24$, $S^2 = 54.6$, $d = 10\%$ of $24 = 2.4$ and $(1-\alpha) = 0.95$ or $\alpha = 0.05$.

We know that

$$P\left[\frac{|\bar{y}-\overline{Y}|}{S\sqrt{\frac{1}{n}-\frac{1}{N}}} \geq 1.96\right] = 0.05 \text{ (assuming sample size large)}$$

Thus, $$d = 1.96\, S\sqrt{\frac{1}{n}-\frac{1}{N}}$$

or, $$d^2 = (1.96)^2\, S^2\left(\frac{1}{n}-\frac{1}{N}\right)$$

Case I : SRSWR

In this case since fpc can be ignored, we have

$$d^2 = \frac{(1.96)^2 S^2}{n}$$

$\therefore$ The sample size, $n = \frac{(1.96)^2 S^2}{d^2} = \frac{(1.96)^2\, 54.6}{(2.4)^2} = 36.4 \cong 37$

6.17 STRATIFIED SAMPLING

6.17.1 Intrduction

In Theorem 6.3, we proved that under SRSWOR from a finite population,

$$\text{Var}(\bar{y}) = \frac{N-n}{N} \cdot \frac{S^2}{n} = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

This shows that in SRSWOR, the precision of an estimtor of the population mean and hence the population total depends not only on the sample size 'n' but also on the variability or heterogeneity among the units of the population, S^2 [because, Var $(\bar{y})$ is directly proportional to S^2 and inversely proportional to n]. This indicates that the Var $(\bar{y})$ can be reduced and hence the precision can be increased either by increasing the sample size n or by reducing the heterogeneity of the population units S^2 or by both. Apart from increasing the sample size, one possible way to estimate the population mean with greater precision is to divide the population into several groups each consisting of more homogeneous units than the population taken as a whole and then to draw simple random samples of pre-determined sizes from each one of the groups. The groups, into which the population is divided, are called strata and the whole procedure of dividing the population into strata and then drawing a random sample from each one of the strata is called stratified random sampling. For example, to estimate the average per capita income of the people living in Bhubaneswar, the characteristic under study would be the income of individuals. Since income varies from individual to individual and there may be wide difference among the income of some than those of the others, the population of Bhubaneswar may be characterised as rich class, middle class and poor class income group of persons. Each of the groups is called a stratum and more than one group strata. From each of these of the same size drawn from the whole population.

In simple random sampling there is no guarantee that all the segments of the population will be adequately represented in the sample. Stratified sampling, on the other hand, enables one to draw a sample representing different segments of the heterogeneous population to any desired extent. Therefore, stratification is frequently used in designing sample surveys. Of course, stratified sampling pre-supposes the knowledge of strata sizes and the availability of a suitable frame for selecting samples from each stratum. Stratified sampling, apart from increasing the precision of the estimated popluation mean, is also commonly used to provide estimates of the total or the mean for the different sub-divisions constituting the population.

Remarks (1): In stratified sampling the two points, viz.,

(i) proper classification of the population into various strata, and

(ii) a suitable sample size from each stratum

are equally important. If the stratification is faulty, the error accrued cannot be compensated even by taking a large sample.

(2) The criterion which enables one to classify various sampling units into different strata is termed as stratifying factor. Some of the commonly used stratifying factors are age, sex, education, income level, geographical area and so on, A stratifying factor is called effective if it divides the given population into different strata so that (i) units (individuals) within each startum are as homogeneous as possible and (ii) the strata means are as wide as possible. It may further be noted that the strata are mutually exclusive i.e., each unit of the population belongs to one and only one of the strata.

(2) The stratified random sampling procedure provides more precise estimates than simple random sampling. Moreover, through this procedure, bias due to selection of unrepresentative samples can be avoided to a great extent.

(3) In simple random sampling procedure, while dealing with heterogeneous population, a fairly large sample size would be needed for proper representation of the populaion. However, in stratified random sampling, this objective cɔn be achieved with a smaller sample size. Thus, stratified random sampling saves a lot of time, money and other resources for data collection.

Disadvantages :

(1) The main disadvantage of stratified random sampling procedure is that, a detailed knowledge of the distribution of the characteristics in the population is needed before hand. If homogeneous groups cannot be identified properly, it is better to go for simple random sampling procedure since improper stratification leads to serious and non-compensatory errors.

(2) The other disadvantage of stratified random sampling is the preparation of lists (one for each stratum) of the population units in various strata. As the list of the popula:ion units is usually not available for each characteristic, the preparation of such lists may be a very difficult task.

6.17.3 Notations

Assume that a population of size N is divided into k strata on the basis of certain characteristic and that samples are taken from each stratum by simple random sampling without replacement procedure, unless otherwise specified. Let i= 1, 2, ..., k denote the stratum and j = 1, 2, ..., N_i denote the sampling unit within the stratum.

2	Y_{21} Y_{22} Y_{23} Y_{2j} Y_{2N_2}	N_2	$\sum_{j=1} Y_{2j}$
i	Y_{i1} Y_{i2} Y_{i3} Y_{ij} Y_{iN_i}	N_i	$\sum_{j=1}^{N_i} Y_{ij}$
k	Y_{k1} Y_{k2} Y_{k3} Y_{kj} Y_{kN_k}	N_k	$\sum_{j=k}^{N_k} Y_{kj}$
	Total	N	$N = \sum_{i=1}^{k} N_i$

Let $y_{i1}, y_{i2}, \ldots\ldots, y_{in_i}$ be a simple random sample of size n_i taken from the ith stratum such that $\sum_{i=1}^{k} n_i = n$, the total sample size.

We adopt the following symbols for the ith stratum; $i = 1, 2, \ldots, k$.

N_i = total number of units

n_i = number of units in the sample

$W_i = N_i / N$ the stratum weight

$f_i = n_i / N_i$ the sampling fraction

Y_{ij} = the value of the characteristic under study for the jth unit; $j = 1, 2, \ldots\ldots, N_i$

y_{ij} = the value of the j th sampled unit.

$\overline{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}$, the stratum mean

$s_i^2 = \frac{\quad}{(n_i-1)}\sum_{j=1}(y_{ij}-\bar{y}_i)^2$, the sample mean square

6.17.4 Estimation of the Population Mean and its Variance

Denoting the population mean by $\bar{Y}$, we have,

$$\bar{Y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{N_i} Y_{ij}}{N} = \frac{\sum_{i=1}^{k} N_i\bar{Y}_i}{N} = \sum_{i=1}^{k} W_i\bar{Y}_i \qquad \text{...... (6.37)}$$

We define an estimator $\bar{y}_{st}$ (st denotes stratified random sampling) for the population mean $\bar{Y}$ as

$$\bar{y}_{st} = \sum_{i=1}^{k} W_i\,\bar{y}_i = \sum_{i=1}^{k} N_i\,\bar{y}_i\,/\,N \qquad \text{...... (6.38)}$$

It may be noted that the estimator $\bar{y}_{st}$ is different from overall sample mean denoted by

$$\bar{y} = \sum_{i=1}^{k} n_i\,\bar{y}_i\,/\,n \qquad \text{...... (3.39)}$$

It may also be noted that $\bar{y}$ coincides with $\bar{y}_{st}$ when the sampling fractions are the same for all the strata, i.e.,(n_i/N_i)'s are all equal.

Theorem 6.6 $\bar{y}_{st} = \sum_{i=1}^{k} W_i\,\bar{y}_i$ is an unbiased estimator of $\bar{Y}$ and its variance is given by,

$$\text{Var}\,(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2\,\frac{(N_i-n_i)}{N_i}\,\frac{S_i^2}{n_i} \qquad \text{...... (6.40)}$$

Proof. Since sampling within each stratum is simple random sampling without replacement, the sample mean $\bar{y}_i$ in the ith stratum is an unbiased estimator of the population mean $\bar{Y}_i$ in the ith stratum i.e., This shows that $\bar{y}_{st}$ is an unbiased estimator of $\bar{Y}$. To obtain the variance of the estimator, we note that sampling is done independently in each stratum and, therefore,

$$\text{Var}(\bar{y}_{st}) = \text{Var}\left(\sum_{i=1}^{k} \frac{N_i\,\bar{y}_i}{N}\right) = \text{Var}\left(\sum_{i=1}^{k} W_i\,\bar{y}_i\right)$$

$$= \sum_{i=1}^{k} W_i^2\ \text{Var}\ (\bar{y}_i)$$

$$= \sum_{i=1}^{k} W_i^2 \left(\frac{N_i - n_i}{N_i}\right) \frac{S_i^2}{n_i} \qquad (6.40a)$$

Thus, we see that Var($\bar{y}_{st}$) depends on S_i^2, the heterogeneity within the ith stratum. If S_i^2 (for all i =1, 2, ..., k) are small, i.e., strata are homogeneous within themselves, then stratified sampling would provide estimates with greater precision.

Corollary 6.5 If $\hat{Y}_{st} = N\,\bar{y}_{st}$ is the estimator of the population total Y, then $\hat{Y}_{st}$ is an unbiased estimator of Y and its sampling variance is given by

$$\text{Var}\ (\hat{Y}_{st}) = \text{Var}\ (N\ \bar{y}_{st})$$

$$= N^2\ \text{Var}\ (\bar{y}_{st})$$

$$= N^2 \sum_{i=1}^{k}\ W_i^2 \frac{(N_i - n_i)}{N_i} \times \frac{S_i^2}{n_i}$$

$$= N^2 \sum_{i=1}^{k} \frac{N_i^2}{N^2} \frac{(N_i - n_i)}{N_i} \times \frac{S_i^2}{n_i}$$

$$= \sum_{i=1}^{k} N_i\ (N_i - n_i)\ \frac{S_i^2}{n_i}$$

$\frac{(1-f)}{n} \sum_{i=1}^{k} W_i S_i^2$, for

$$\text{Var}(\bar{y}_{st}) = \frac{(N-n)}{N} \sum_{i=1}^{k} N_i S_i^2 / nN \quad \text{(using the above condition)}$$

$$= \frac{(1-f)}{n} \sum_{i=1}^{k} W_i S_i^2 \qquad \text{..... (6.42)}$$

Corollary 6.7 Usually, the S_i^2 values are not available. Since simple random samples are taken from each stratum,

$$E(s_i^2) = S_i^2;\ i = 1, 2, \ldots, k,$$

where $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$, the sample mean square in the ith stratum.

This enables one to get an unbiased estimator of the variance of $\bar{y}_{st}$. This unbiased estimator is given by

$$\text{Est Var}(\bar{y}_{st}) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 s_i^2 \qquad \text{..... (6.43)}$$

$$= \frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) \frac{s_i^2}{n_i}.$$

Corollary 6.8 In stratified random sampling if the sampling fraction (n_i / N_i)'s for all strata are negligible, then (6.40a) and (6.43), respectively, reduce to

$$\text{Var}(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{N_i^2 S_i^2}{N^2 n_i} \qquad \text{.... (6.44)}$$

(ii) the variability within the stratum, and

(iii) the cost per sampling unit in the stratum.

A good allocation is one where maximum precision is obtained with minimum resources, or in other words, the criterion for allocation is to minimize the budget for a given variance or minimize the variance for a fixed budget, thus making the most effective use of the available resources.

There are two methods of allocation of sample sizes to different strata in stratified random sampling procedure, viz.,

(i) Proportional allocation, and

(ii) Optimum allocation.

6.17.6 Proportional Allocation

The method of proportional allocation, originally proposed by Bowley (1926), is very common in practice because of its simplicity. When no other information about the strata, except their sizes N_i ($i = 1, 2, \ldots, k$) is available, the allocation of a sample of size n_i, out of a total given sample of size n is done for the ith stratum in proportion to its size, i.e., for the ith stratum, it is n_i / N_i, where $\sum_{i=1}^{k} n_i = n$ and $\sum_{i=1}^{k} N_i = N$.

Allocation of sample sizes is called proportional if for all strata, the sampling fractions are equal i.e. $(n_i / N_i) = f_i = f$ for all $i = 1, 2, \ldots, k$. Thus, we have

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots\ldots = \frac{n_k}{N_k}$$

Applying the principle of ratio and proportion, we get

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots\ldots = \frac{n_k}{N_k} = \frac{n_1 + n_2 + \ldots + n_k}{N_1 + N_2 + \ldots + N_k} = \frac{n}{N} = \text{Ç, (constant)}$$

or, $$n_i = \frac{n}{N} . N_i ; \quad i = 1, 2, ..., k \qquad ......(6.46)$$

or, $$n_i \propto N_i$$

With proportional allocation, the sampling fraction in all strata being equal, it gives a self weighting sample. If numerous estimates have to be made, a self weighting system saves time and results in gain in precision.

The expression for Var($\overline{y}_{st}$) under proportional allocation is the same that has been discussed in Corollary 6.6, i.e.,

$$Var(\overline{y}_{st}) = \frac{(N-n)}{N} \sum_{i=1}^{k} W_i S_i^2 / n$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i S_i^2 \qquad ...... (6.47)$$

Note : If different strata have similar variances of the characteristic being measured, then the statistical efficiency will be the highest under proportional allocation. This is evident from the fact that, since the sampling units in each stratum are homogene-ous, S_i^{2}'s would naturally be smaller. When all the S_i^{2}'s are equal; S^2 say, we get,

$$Var(\overline{y}_{st})_{prop} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i S^2$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \sum_{i=1}^{k} W_i$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) S^2.$$

6.17.7 Optimum Allocation

The expression for $\mathrm{Var}(\overline{y}_{st})$, given in (6.40) shows that the precision of the estimator of the population mean based on stratified samples, depends on n_i which can be fixed in advance by the surveyor. The guiding principle in the determination of n_i's is to choose them in such a manner that (a) the variance of the estimate is minimised (or the precision is maximised) for fixed sample size 'n' and fixed cost 'C' and (b) the total cost is minimised for (fixed) given precision. This would make the most effective use of the resources available. The allocation of the sample sizes to the different strata made in accordance with this principle is called the optimum allocation. This concept of optimum allocation can be stated as : optimum allocation aims at allocation of n_i's such that

(i) $\mathrm{Var}(\overline{y}_{st})$ is minimum for fixed n

(ii) $\mathrm{Var}(\overline{y}_{st})$ is minimum when the total cost C is fixed and

(iii) total cost C is minimum for fixed precision.

Cost Function :

In any sample survey, the value of information on the experimental units must always be balanced against the cost of obtaining it. In stratified random sampling, since the sampling units vary in quality for different strata, it may cost more to obtain information about a sample in one stratum than in another. Keeping in mind the above fact, we define, in the simplest form, the cost function C in stratified sampling as

$$C = C_0 + \sum_{i=1}^{k} c_i n_i \qquad \text{..... (6.48)}$$

where 'C_0' is the overhead cost and c_i is the cost per unit in the ith stratum.

Theorem 6.7 : In stratified random sampling, the variance of the estimated mean $\overline{y}_{st}$ is minimum for a fixed total size of the sample n if $n_i \propto N_i S_i$. In other words, $\mathrm{Var}(\overline{y}_{st})$ is minimum for fixed n if

$$n_i \propto N_i S_i$$

Proof : We have to minimise

$$\mathrm{Var}(\overline{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i}$$

subject to the restriction

$$\sum_{i=1}^{k} n_i = n \text{ (fixed)}$$

Using the technique of Lagrange multiplier, this is equivalent to minimising

$$\phi = Var(\overline{y}_{st}) + \lambda \left(\sum_{i=1}^{k} n_i - n\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i} + \lambda \left(\sum_{i=1}^{k} n_i - n\right)$$

subject to variations in n_i, λ being unknown Lagrange multiplier.

Now, differentiating ϕ w.r.t. n_i and equating to zero, we have

$$\frac{\partial \phi}{\partial n_i} = -\frac{N_i^2 S_i^2}{N^2 n_i^2} + \lambda = 0$$

or, $$n_i = \frac{N_i S_i}{N\sqrt{\lambda}} \qquad \text{..... (6.49)}$$

Further,

$$\frac{\partial^2 \phi}{\partial n_i^2} = \frac{2\, N_i^2\, S_i^2}{N^2 . n_i^3} > 0$$ [as N_i, S_i, N and n_i are all positive quantities]

Thus, the n_i's given by (6.49) provide minimum value for ϕ.

Now, summing (6.49) over i from 1 to k, we get

$$\sum_{i=1}^{k} n_i = n = \sum_{i=1}^{k} N_i S_i / N\sqrt{\lambda}$$

or, $$\sqrt{\lambda} = \sum_{i=1}^{k} N_i S_i / nN \qquad \text{..... (6.50)}$$

Substituting the value of $\sqrt{\lambda}$ from (6.50) in (6.49), we finally get

$$n_i = n\frac{N_i S_i}{\sum_{i=1}^{k} N_i S_i} \qquad \text{..... (6.51)}$$

or, $n_i \propto N_i S_i$. (proved)

This is known as Neyman's formula for optimum allocation. This result suggests that greater the value of N_iS_i for a given stratum, greater is the number of sampling units to be selected from the stratum in order to obtain the most precise estimate of the population mean.

A formula for the variance of $\bar{y}_{st}$ under optimum allocation for a fixed n can be obtained by substituting the value of n_i from (6.51) in the general formula for $Var(\bar{y}_{st})$ as follows :

We know that, $Var(\bar{y}_{st})$ is given by

$$Var\ (\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i}$$

$$= \frac{1}{N^2}\sum_{i=1}^{k} N_i\ (\frac{N_i}{n_i} - 1)S_i^2 .$$

Substituting the value of n_i from (6.51), we get

$$Var(\bar{y}_{st})_{opt} = \frac{1}{N^2}\sum_{i=1}^{k} N_i \left\{\frac{\sum_{i=1}^{k} N_iS_i}{nS_i} - 1\right\}S_i^2$$

$$= \frac{1}{N^2}\left[(\sum_{i=1}^{k}N_iS_i)\frac{(\sum_{i=1}^{k}N_iS_i)}{n} - \sum_{i=1}^{k}N_iS_i^2\right]$$

$$= \frac{1}{N^2}\left[\frac{1}{n}\left(\sum_{i=1}^{k}N_iS_i\right)^2 - \sum_{i=1}^{k}N_iS_i^2\right]$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k}\frac{N_iS_i}{N}\right)^2 - \frac{1}{N}\sum_{i=1}^{k}\frac{N_iS_i^2}{N}$$

Thus, $$Var(\bar{y}_{st})_{opt} = \frac{\left(\sum_{i=1}^{k}W_iS_i\right)^2}{n} - \frac{\sum_{i=1}^{k}W_iS_i^2}{N} \qquad ..... (6.52)$$

There may be difficulty in using this formula as the S_i values are usually unknown. However, the stratum variances S_i^2's may be obtained from past surveys conducted for similar populations or from a specially planned pilot survey.

Theorem 6.8 : In stratified random sampling with a given linear cost function of the form $C = C_0 + \sum_{i=1}^{k} c_i n_i$, where C is the total cost, C_0 is the overhead cost, c_i and n_i are, respectively, the cost per sampling unit and the sample size of the ith stratum,

(i) the variance of the estimated mean $\overline{y}_{st}$ is minimum for a specified cost C,

and (ii) the cost C is minimum for a specified variance of $\overline{y}_{st}$ when n_i is proportional to $W_i S_i / \sqrt{c_i}$, or equivalently, when $n_i \propto N_i S_i / \sqrt{c_i}$.

Proof :

We are given with the cost function :

$$C = C_0 + \sum_{i=1}^{k} c_i n_i$$

or,
$$\sum_{i=1}^{k} c_i n_i = C - C_0 \quad \text{......} (6.53)$$

We know that,

$$\text{Var}(\overline{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) \frac{S_i^2}{n_i}$$

$$= \sum_{i=1}^{k} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2$$

$$= \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i} \quad \text{......} (6.53a)$$

To minimise $\text{Var}(\overline{y}_{st})$ subject to (6.53), we use the method of Lagrange multiplier, Thus, minimising (6.53a) subject to (6.53) is equivalent to minimising.

$$\Psi = \text{Var}(\overline{y}_{st}) + \lambda \left(\sum_{i=1}^{k} c_i n_i - C + C_0\right), \text{where } \lambda \text{ is Lagrange multiplier.}$$

$$= \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i} + \lambda \left(\sum_{i=1}^{k} c_i n_i - C + C_0\right)$$

Differentiating ψ partially with respect to n_i and equating to zero, we have,

$$\frac{\partial \psi}{\partial n_i} = -\frac{W_i^2 S_i^2}{n_i^2} + \lambda c_i = 0$$

$$\Rightarrow \quad n_i^2 = \frac{W_i^2 S_i^2}{\lambda c_i}$$

$$\Rightarrow \quad n_i = \frac{W_i S_i}{\sqrt{\lambda c_i}} \qquad \text{...... (6.54)}$$

$$\text{Now,} \quad \frac{\partial^2 \psi}{\partial n_i^2} = \frac{2W_i^2 S_i^2}{n_i^3} > 0$$

i.e., the value of n_i given by (6.54) minimises ψ.

To find the value of λ, we sum over all strata and find

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} (W_i S_i / \sqrt{\lambda c_i}) \qquad \text{..... (6.55)}$$

$$\text{So,} \quad \sqrt{\lambda} = \frac{1}{n} \sum_{i=1}^{k} (W_i S_i / \sqrt{c_i}) \qquad \text{..... (6.56)}$$

Substituting (6.56) in (6.54), we get

$$n_i = n \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} (W_i S_i / \sqrt{c_i})} \qquad \text{..... (6.57)}$$

Thus, in optimum allocation for a fixed cost

$$n_i \; \alpha \; \frac{W_i S_i}{\sqrt{c_i}} \qquad \text{..... (6.58)}$$

$$\text{Equivalently,} \quad n_i \; \alpha \; \frac{N_i S_i}{\sqrt{c_i}}$$

This theorem leads to the following important conclusion :

A large sample would be required to be drawn from a stratum if

(i) the stratum size N_i is larger,

(ii) the stratum variability (S_i) is larger, and

(iii) the cost per unit is lower in the stratum.

Equation (6.57) gives the value of n_i in terms of n, but the value of n is not known to us as yet. The solution depends on whether the sample is chosen so as to meet a specified total cost C or to give a specified value of the variance of $\overline{y}_{st}$ say, V. If the cost is fixed, we substitute the optimum value of n_i in the cost function (6.48) and solve for n. This gives

$$C = C_0 + \frac{\sum_{i=1}^{k} n c_i N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} N_i S_i / \sqrt{c_i}}$$

or,
$$n = \frac{(C - C_0)\sum_{i=1}^{k} N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} N_i S_i \sqrt{c_i}} \quad \text{..... (6.59)}$$

If $\text{Var}(\overline{y}_{st}) = V$ is fixed, we find

$$V = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i}$$

$$\therefore \quad \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} = V + \sum_{i=1}^{k} W_i \frac{N_i}{N} \frac{S_i^2}{N_i} = V + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2.$$

Substituting the optimum n_i from (6.57) we get,

$$\sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n\, W_i S_i / \sqrt{c_i}} \times \sum_{i=1}^{k} W_i S_i / \sqrt{c_i} = V + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2$$

or,
$$\frac{1}{n} \left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right) \left(\sum_{i=1}^{k} W_i S_i / \sqrt{c_i} \right) = V + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2$$

$$\therefore \quad n = \frac{\left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i}\right)\left(\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}\right)}{V + \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2} \qquad \text{..... (6.60)}$$

Remarks :

If c_i's are the same for all strata, (6.57) will lead to the Neyman allocation. Similarly if c_i's and S_i's do not vary from stratum to stratum, (6.57) will lead to proportional allocation.

6.17.8 Relative Precision of Stratified Random Sampling and Simple Random Sampling

In this section, a comparison is made between simple random sampling and stratified random sampling with proportional and optimum allocations. This comparison shows how the gain due to stratification is achieved. We denote the variances of the estimated means of simple random sampling, stratified random sampling with propotional and optimum allocations by V_{ran}, V_{prop} and V_{opt}, respectively.

Theorem 6.9 : If terms in $1/N_i$ are ignored relative to unity, then

$$V_{opt} \le V_{prop} \le V_{ran}' \qquad \text{..... (6.61)}$$

where V_{opt} denotes the variance under optimum allocation for fixed n, i.e., where $n_i \propto N_i S_i$.

Proof :

We have, respectively, from (6.9), (6.47) and (6.52) :

$$V_{ran} = (1-f)\frac{S^2}{n}, \qquad \text{..... (6.62)}$$

$$V_{prop} = \frac{(1-f)}{n}\sum_{i=1}^{k} W_i S_i^2 = \frac{\sum_{i=1}^{k} W_i S_i^2}{n} - \frac{\sum_{i=1}^{k} W_i S_i^2}{N} \qquad \text{..... (6.63)}$$

and

$$V_{opt} = \frac{\left(\sum_{i=1}^{k} W_i S_i\right)^2}{n} - \frac{\sum_{i=1}^{k} W_i S_i^2}{N} \qquad \text{..... (6.64)}$$

In order to compare (6.62) with (6.63), we first express S^2 in terms of S_i^2, i.e,

$$
\begin{aligned}
(N-1)\,S^2 &= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\overline{Y})^2\\
&= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\overline{Y}_i+\overline{Y}_i-\overline{Y})^2\\
&= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\overline{Y}_i)^2+\sum_{i=1}^{k}N_i(\overline{Y}_i-\overline{Y})^2 \quad \text{(the cross product terms vanish when sum is taken over j)}\\
&= \sum_{i=1}^{k}(N_i-1)\,S_i^2+\sum_{i=1}^{k}N_i(\overline{Y}_i-\overline{Y})^2 \qquad \text{......(6.65)}
\end{aligned}
$$

If terms containing $1/N_i$ and hence $1/N$ are neglible, (6.65) reduces to

$$S^2=\sum_{i=1}^{k}W_iS_i^2+\sum_{i=1}^{k}W_i(\overline{Y}_i-\overline{Y})^2 \qquad \text{...... (6.66)}$$

Hence, $$V_{ran}=V_{prop}+\frac{(1-f)}{n}\sum_{i=1}^{k}W_i(\overline{Y}_i-\overline{Y})^2 \qquad \text{...... (6.67)}$$

or, $$V_{ran}-V_{prop}=\text{a non-nagative quantity}\geq 0$$

Thus, $$V_{prop}\leq V_{ran}. \qquad \text{...... (6.68)}$$

From (6.67), it is clear that the larger the difference in the stratum means, the greater is the gain in precision with proportional allocation over simple random sampling.

Similarly,

$$
\begin{aligned}
V_{prop}-V_{opt} &= \frac{1}{n}\left[\sum_{i=1}^{k}W_iS_i^2-(\Sigma\, W_iS_i)^2\right]\\
&= \frac{1}{n}\left[\sum_{i=1}^{k}W_i(S_i-\overline{S})^2\right] \qquad \text{...... (6.69)}
\end{aligned}
$$

where $\bar{S} = \sum_{i=1}^{k} W_iS_i = \frac{1}{N}\sum_{i=1}^{k} N_iS_i$ is a weighted mean of the stratum standard deviations, the weights being equal to stratum sizes.

Since the R.H.S. of (6.69) is non-nagative, $V_{prop} - V_{opt} \geq 0$

$$\text{i.e.,}\quad V_{opt} \leq V_{prop} \qquad \text{..... (6.70)}$$

Combining (6.68) and (6.70), we arrive at the required result, i.e.,

$$V_{opt} \leq V_{prop} \leq V_{ran}.$$

From (6.69) we conclude that, greater the difference between the stratum standard deviations, greater is the gain in precision of Neymans allocation over proportional allocation.

Again, from (6.67) and (6.69), with terms in $1/N_i$ negligible,

$$V_{ran} = V_{opt} + \frac{1}{n}\sum_{i=1}^{k} W_i(S_i - \bar{S})^2 + \frac{(1-f)}{n}\sum_{i=1}^{k} W_i(\bar{Y}_i - \bar{Y})^2$$

$$= V_{opt} + \frac{1}{n}\sum_{i=1}^{k} W_i(S_i - \bar{S})^2 + \frac{1}{n}\Sigma W_i(\bar{Y}_i - \bar{Y})^2 \qquad \text{..... (6.71)}$$

or, $V_{ran} - V_{opt}$ = a non-nagative quantity ≥ 0

$$\text{i.e.,}\quad V_{ran} \geq V_{opt} \qquad \text{..... (6.72)}$$

Thus, we observe that

$$V_{prop} \leq V_{ran},$$

$$V_{opt} \leq V_{prop}$$

$$\text{and}\quad V_{opt} \leq V_{ran}$$

$$\text{Hence,}\quad V_{opt} \leq V_{prop} \leq V_{ran}$$

We observe from (6.71) that as we change from unstratified simple random sampling to stratified random sampling with Neyman's allocation, the gain in precision of the estimators results from two factors, viz,

(i) the elimination of the differences among the stratum means, and

(ii) the elimination of the differences among the stratum standard deviations.

Example 6.10

A sample of 200 units is to be drawn from a population consisting of 1000 units. The population is divided into four homogeneous groups on the basis of location and income as follows :

High income - urban = 200

Low income - urban = 400

High income - rural = 100

Low income - rural = 300

Determine the sizes of samples to be drawn from different strata by proportional allocation method.

Solution :-

Here, we are given $N = 1000$, $N_1 = 200$, $N_2 = 400$, $N_3 = 100$, $N_4 = 300$ and $n = 200$

The proportion of sample size to the population size in each stratum should be

$$\frac{n}{N} = \frac{200}{1000} = 0.2$$

Let n_i be the sample size from stratum of size N_i; $i = 1, 2, 3, 4$.

$$\therefore \quad \frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \frac{n_4}{N_4} = 0.2$$

which gives

$$n_1 = 0.2 \times N_1 = 0.2 \times 200 = 40,$$

$$n_2 = 0.2 \times N_2 = 0.2 \times 400 = 80,$$

$$n_3 = 0.2 \times N_3 = 0.2 \times 100 = 20,$$

and $\quad n_4 = 0.2 \times N_4 = 0.2 \times 300 = 60.$

Example 6.11

For the Example 6.10, the following additional information is given :

Stratum	variance (S^2)
High income - urban	6.5
Low income - urban	2.5
High income - rural	4.5
Low income - rural	2.0

Use the disproportional stratified sampling procedure (optimum allocation) to choose a sample of size 200 from the four strata.

Solution: -

The stratum sample size using disproportional stratified sampling (Neyman's optimum alloction) procedure is given by

$$n_i = n\frac{N_iS_i}{\sum N_iS_i}, i=1,\ 2,\ 3,\ 4$$

Table 6.7

Computation of sample size according to Neyman's Allocation.

Stratum	Stratum Size (N_i)	Stratum Variance (S_i^2)	Stratum Standard deviation(S_i)	N_iS_i	Sample size $n_i = \frac{nN_iS_i}{\sum N_iS_i}$
High income - urban	200	6.5	2.5	500	57
Low income - urban	400	2.5	1.6	640	72
High income - rural	100	4.5	2.1	210	24
Low income - rural	300	2.0	1.4	420	47
Total	1000			1770	200

Thus, the sample sizes in the four strata are found to be 57, 72, 24 and 47, respectively.

Example 6.12

Given the following data from a stratified sample, calculate the estimates of the population mean and the standard error of the estimate :

Stratum	N_i	n_i	$\bar{y}_i$	s_i^2
1	10	4	3.4	1.50
2	15	5	2.8	0.90
3	25	11	4.4	1.25

Solution : -

Table - 6.8

The following table is prepared for computational convenience.

Stra-tum No	N_i	n_i	$\bar{y}_i$	s_i^2	$N_i\, n_i$	$N_i\, \bar{y}_i$	(N_i/N)	$N_i - n_i$	$(N_i - n_i\, N_i\, n_i)$	$(N_i/N)^2(N_i/n_i)s_i^2$ $\overline{N_i\, n_i}$
					(2)x(3)	(2)x(4)	$\{(2) \div 50\}^2$	(2) − (3)	(9)) ÷ (6)	(8)x(10)x(15)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	10	4	3.4	1.50	40	34.0	0.04	6	0.1500	0.0090
2	15	5	2.8	0,90	75	42.0	0.09	10	0.1333	0.0108
3	25	11	4.4	1.25	275	110.0	0.25	14	0.0509	0.0159
Total	50					186.0				0.0357

$N = \sum N_i = 50$

The estimate of the population mean is

$$\bar{y}_{st} = \frac{\sum N_i \bar{y}_i}{N} = \frac{186}{50} = 3.72$$

The estimate of standard error of $\bar{y}_{st}$ is

$$\text{Est SE}(\bar{y}_{st}) = \sqrt{\sum \left(\frac{N_i}{N}\right)^2 \frac{N_i - n_i}{N_i n_i} S_i^2} = \sqrt{0.357}$$

$$= 0.1889 \approx 0.19$$

Example 6.13

A sample of 30 students is drawn from a population consisting of 300 students belonging to two colleges A and B. The means and standard deviations of their marks are given below :

College	Total No. of Students (N_i)	Mean $\bar{Y}_i$	Standard Deviation S_i
A	200	50	10
B	100	80	15

(i) Determine the sample sizes by the proportional and optimum allocation methods for a total sample of 30 students.

(ii) Compute the standard error of the estimte of mean marks based on the sample of 30 students by

(a) unstratified random sampling

(b) stratified random sampling under proportional allocation, and

(c) stratfied random sampling under optimum allocation

(iii) compare the efficiencies for the above methods of sampling.

Solution :-

Table 6.9

(i) Computation of Sample Sizes by Proportional and Optimum Allocations

Stratum	N_i	S_i	N_iS_i	Proportional allocation nN_i / N	Optimum allocation $\frac{nN_iS_i}{\Sigma N_iS_i}$
1	200	10	2000	20	17
2	100	15	1500	10	13
Total	300		3500	30	30

The sample sizes allocated to the strata have been calculated as indicated under the last two columns of the Table 6.9.

Table 6.10

(ii) Computation of Standard Error of Estimate

N_i	W_i	$\bar{Y}_i$	S_i	W_iS_i	$W_iS_i^2$	$(\bar{Y}_i - \bar{Y})^2$	$W_i(\bar{Y}_i - \bar{Y})^2$
200	0.67	50	10	6.7	67	100	67
100	0.33	80	15	5.0	74	400	132
N=300	1.0	–	–	11.7	141	–	199

(a) In case of simple random sampling, the variance of the estimate of the population mean is

$$V(\bar{y}) = \frac{N-n}{Nn} S^2$$

$$= \frac{N-n}{Nn}\left[\sum_i \frac{N_i - 1}{N-1} S_i^2 + \sum_i \frac{N_i}{N-1}(\bar{Y}_i - \bar{Y})^2\right]$$

$$\cong \frac{N-n}{Nn}\left[\sum_i W_iS_i^2 + \sum W_i(\bar{Y}_i - \bar{Y})^2\right] \quad \text{(for large } N_i \text{ and N)}$$

$$= \frac{300-30}{300 \times 30}\left[141 + 199\right]$$

$$= 10.2$$

Thus, the standard error of the estimate is

$$SE(\bar{y}) = \sqrt{Var(\bar{y})} = \sqrt{10.2} = 3.2 \text{ marks}.$$

(b) When the sample is drawn from each stratum following the method of proportional allocation, we have

$$Var(\bar{y}_{st})_{prop} = \frac{N-n}{Nn} . \sum W_iS_i^2$$

$$= \frac{270}{300 \times 30} \times 141$$
$$= 4.23$$

Thus, the standard error of $\overline{y}_{st}$ with propotional allocation is

$$\text{S.E. } (\overline{y}_{st})_{prop} = \sqrt{4.23} = 2.05 \text{ marks .}$$

(c) The variance of $\overline{y}_{st}$ under optimum allocation is computed as

$$\text{Var } (\overline{y}_{st})_{opt} = \frac{1}{n} \left(\Sigma W_i S_i\right)^2 - \frac{1}{N} \left(\Sigma W_i S_i^2\right)$$

$$= \frac{1}{30} \times (11.7)^2 - \frac{1}{300} \times 141$$
$$= 4.56 - 0.47$$
$$= 4.09.$$

Hence , $\text{SE } (\overline{y}_{st})_{opt} = \sqrt{4.09} = 2.02$ marks.

(iii) Gain in efficiency of the stratified random sampling with proportional allocation compared to simple random sampling is given by

$$\frac{\text{Var } (\overline{y}) - \text{Var } (\overline{y}_{st})_{prop}}{\text{Var } (\overline{y}_{st})_{prop}} \times 100$$

$$= \frac{10.2 - 4.23}{4.23} \times 100 = 140\%.$$

Gain in efficiency of the stratified random sampling with optimum allocation compared to simple random sampling is computed as

$$\frac{\text{Var } (\overline{y}) - \text{Var } (\overline{y}_{st})_{opt}}{\text{Var } (\overline{y}_{st})_{opt}} \times 100 = \frac{10.2 - 4.09}{4.09} \times 100 = 150\%.$$

Gain in efficiency of the stratified random sampling with optimum allocation compared to stratified random sampling with proportional allocation is as follows :

$$\frac{\text{Var } (\overline{y})_{prop} - \text{Var } (\overline{y}_{st})_{opt}}{\text{Var } (\overline{y}_{st})_{opt}} \times 100 = \frac{4.23 - 4.09}{4.09} \times 100 = 3.4\%.$$

***

EXERCISES - 6.1

1. What is a sample survey ? What are the advantages of sample survey over complete census ?
2. What are the basic principles of sample survey ? Discuss.
3. What are the principal steps followed in the planning and execution of a large scale sample survey ? Explain.
4. Distinguish between sampling error and non-sampling error. Explain the various sources of sampling error.
5. What are the different sources of errors in a sample survey ? Describe how these errors can be controlled.
6. What is non-sampling error ? Explain briefly the sources of non-sampling error.
7. Define (i) sampling unit, (ii) sampling frame, (iii) finite population correction factor and (iv) standard error of an estimate.
8. Distinguish between a questionnaire and a schedule. Discuss how they are used in the collection of data.
9. Explain what you understand by probability sampling and non-probability sampling. What are their relative advantages and disadvantages ?
10. Describe the lottery method for drawing a random sample from a finite population. Discuss the merits and demerits of this method.
11. What are random numbers ? Explain by means of an illustration how these are used to select a simple random sample without replacement.
12. Distinguish between simple random sampling with replacement and simple random sampling without replacement. Show that, in both the cases, the sample mean provides an unbiased estimate for the population mean.
13. Prove that, in SRSWOR, the probability of selecting a specified unit of the population at any given draw is equal to the prabability of selecting it at the first draw.

14. In SRSWOR, find the following probabilities :

    (i) a specified unit of the population is included in the sample, and

    (ii) two specified units of the population are selected in the sample.

15 Prove that, in SRSWOR, the sample mean square is an unbised estimate of the population mean square.

16. Show that, in SRSWOR, the variance of the sample mean is given by

$$\text{Var}\ (\bar{y}) = \frac{N-n}{N}.\frac{S^2}{n}$$

What is the standard error (S. E) of $\bar{y}$ and how will you estimate it ?

17. Prove that SRSWOR provides a more efficient estimator of the population mean relative to SRSWR.

18. What is simple random sampling for attributes ? In a population of size N, the number of units possessing a certain characteristic is 'A' and in a simple random sample of size n from it, the number of units possessing that characteristic is 'a'.

If $P = \frac{A}{N}$ and $p = \frac{a}{n}$; $Q = 1 - P$ and $q = 1 - p$, then prove that

(i) p is an unbiased estimate of P,

(ii) $\text{Var}\ (p) = \frac{N-n}{N-1}\frac{PQ}{n}$, and

(iii) an unbiased estimate of Var (p), $v\ (p) = \frac{N-n}{n-1}.\frac{pq}{N}$

19. Discuss the method of determining the sample size in case of simple random sampling without replacement so as to meet the desired margin of error and confidence, stating the assumptions made.

20. A simple random sample without replacement of size, n = 20 was drawn from a village in which there are N = 200 households. It was found that amongst the sampled households there were only 8 households each possessing a TV. Estimate th total number of households in the village possessing TV sets and calculate the standard error of the estimate.

21. In selecting 2 units with SRSWOR from a population having 4 units with values 5, 8, 12, 15, show that the sample mean is an unbiased estimator of the population mean by enumerating all possible samples. Caculate its sampling variance and verify that it agrees with the formula for the variance of sample mean.

22. What is stratified random sampling ? When will you use stratified random sampling ? Describe the advantages of stratified random sampling with illustrations.

23. Explain the procedure of stratified random sampling. Under what conditions is stratified random sampling preferred to simple random sampling and why ?

24. Explain the purpose of stratification in sample surveys. Propose an unbiased estimate of the population mean by the method of stratified random sampling and obtain the expression for its variance.

25. Discuss Neyman's optimum allocation principle in stratified random sampling.

26. Compare the efficiencies of the Neyman and proportional allocations with that of an unstratified random sample of the same size.

27. With a cost function of the type $C = a + \sum_i c_i n_i$, prove that the variance of the estimated mean $\overline{y}_{st}$ is minimum when n_i is proportional to $N_i S_i / \sqrt{c_i}$

28. A sample of 30 units is to be drawn from a population consisting of 300 students belonging to two colleges A and B. The mean and standard deviation of their marks are given below :

College	Total number of Students (N_i)	Mean ($\overline{Y}_i$)	Standard deviation (S_i)
Collage A	180	70	10
Collage B	120	50	15

How would you draw the sample using optimum allocation scheme ? Obtain the variance of the estimate of the population mean under this scheme.

29. A population of size 500 is divided into 3 strata. Their sizes and standard deviations are as follows :

		Strata	
No	I	II	III
Size	100	150	250
SD	5	6	10

A stratified random sample of size 75 is drawn from the population. Determine the sizes of samples from the three strata under (i) proportional allocation, (ii) Neyman's allocation. Explain why the stratum sample sizes are different under these schemes.

30. A stratified sample of 30 units gives the following estimated stratum means and variances :

Stratum No	N_i	n_i	$\bar{y}_i$	s_i^2
1	20	3	15	9
2	25	5	25	16
3	40	12	10	36
4	35	10	20	36

(a) Estimate the variance within the whole population,

(b) Verify that the existing allocation is optimal for the 4 strata, and

(c) Estimate the sampling variance of the estimated population mean for the above allocation and for a random sample of size 30 drawn without stratification and comment.

EXERCISE - 6.2

1.

(i) Define a finite population and a sample with examples.

(ii) What is sampling ?

(iii) What is random sampling ?

(iv) What is non-random sampling ?

(v) Define a random sample.

(vi) What is sampling unit ?

(vii) What is a sampling frame ? Give an example.

(viii) Distinguish between a questionnaire and a schedule.

(ix) What is complete census ? Where is it recommended ?

(x) Why is sampling necessary ?

(xi) Distinguish between sampling error and non-sampling error.

(xii) What is response error ?

(xiii) How are sampling error and sample size related ?

(xiv) Give two situations where sampling is indispensable.

(xv) What is a pilot sample survey ?

2.

(i) Distinguish between SRSWOR and SRSWR.

(ii) What is a Random Number Table ?

(iii) Distinguish between a parameter and a statistic.

(iv) Give the difference between an estimator and an estimate.

(v) Define unbiasedness of an estimate.

(vi) Define precision of an estimate.

(vii) Distinguish between bias and error of an estimator.

(viii) What are the criteria of a good estimator ?

(ix) What is sampling fraction ?

(x) Define finite population correction factor.

(xi) What is the unbiased estimate for population total in SRS ?

(xii) Give the expression for variance of the estimate in SRSWOR.

(xiii) What is standard error of an estimate ?

(xiv) Compare the efficiencies of SRSWOR and SRSWR.

(xv) What is simple random samping for attributes ?

3. (i) Define stratified random sampling.

(ii) What are the reasons for stratification ?

(iii) What is a stratifying factor ?

(iv) Give some examples of stratifying factor.

(v) What is allocation problem in stratified sampling ?

(vi) What are the different types of allocation ?

(vii) What is proportional allocation ?

(viii) Define uniform sampling fraction ?

(ix) What is optimum allocation ?

(x) What is Neyman allocation ?

(xi) Define cost function in stratified random sampling.

(xii) How would you allocate uniformly a sample of size 40 among 4 strata having 20, 60, 70, 50 units ?

(xiii) How do you estimate population total in stratified sampling?

(xiv) Write the expression for the variance of the estimate in stratified sampling.

(xv) Does stratification always result in gain in precision?

ANSWERS

Exercises - 6.1

20. 80, 0.11, 21. Var $(\bar{y})$= 4.8, 28. 15. 15, Var $(\bar{y}_{st})_{opt}$ = 4.3

29. Proportional allocation : 15, 22, 38

Optimum allocation : 10, 17, 48

30. (a) 0.71 (c) 0.60 and 0. 87. Simple random sampling fares poorly as compared to stratified random sampling.

***

MODEL QUESTIONS

1. **Choose the correct answer from the given alternatives :**

(a) A sample consists of :

(i) All the units of the population

(ii) 10 percent of all the units of the population

(iii) 50 percent of all the units of the population

(iv) any fraction of the population

(b) A population consisting of all the out comes of a coin tossing experiment until the first head appears is known as :

(i) finite population

(ii) real population

(iii) hypothetical population

(iv) infinite population

(c) The number of possible samples of size n from a population of size N when SRSWOR is adopted is :

(i) $^{N}C_{n}$

(ii) N^{n}

(iii) n^{N}

(iv) n!

(d) The number of possible samples of size n from a population of size N when SRSWR is followed is :

(i) $^{N}C_{n}$

(ii) N^{n}

(iii) n^{N}

(iv) n!

(e) A sampling frame is :

(i) a list of all units in a sample

(ii) a list of random numbers

(iii) a list of all sampling units of a population

(iv) a list of households

(f) A simple random sample can be drawn with the help of :

(i) random numbers table

(ii) lottery method

(iii) roulette method

(iv) all the above

(g) In simple random sampling, the probability of including a specified unit of the population of size N in the sample of size n is :

(i) $\frac{1}{N}$

(ii) $\frac{1}{n}$

(iii) $\frac{n}{N}$

(iv) $\frac{N}{n}$

(h) Stratified random sampling is recommended when :

(i) The population is homogeneous

(ii) The population is heterogeneous

(iii) The population is hypothetical

(iv) The population is infinite

(i) Under equal allocation in stratified sampling, the sample from each stratum is :

(i) Proportional to stratum size

(ii) of equal size

(iii) in proportion to the per unit cost of survey in the stratum

(iv) all the above

(j) Stratified sampling comes under the category of :

(i) unrestricted sampling

(ii) subjective sampling

(iii) purposive sampling

(iv) resticted sampling

2. **(a) Fill in the blanks :**

(i) A measurable function of all the values constituting a population is________.

(ii) Sampling error and sample size are ________ related.

(iii) Between SRSWOR and SRSWR, the more efficient sampling design is __________.

(iv) If fpc is ignored, then var ($\bar{y}$) in SRSWOR is given by __________.

(v) Sample under proportional allocation is a __________ sample.

(b) Write True or False :

(i) "Greater scope" is one of the advantages of sampling over census.

(ii) There is no difference between a questionnaire and a schedule corresponding to a sample survey.

(iii) Non-sampling errors are present only in census surveys.

(iv) Simple random sampling requires an up-to-date frame.

(v) Neyman's optimum allocation suggests that greater the value of the product N_iS_i, the smaller is the number of units to be sampled from the ith stratum.

3. **Answer the following questions in one or two sentences :**

(a) Give an example where sampling is the only course of action.

(b) Distinguish between a parameter and an estimator.

(c) What is meant by substitution ?

(d) Give the difference between a questionnaire and a schedule.

(e) $\frac{N-n}{N}$ represents a currection for the finite population. What does it mean ?

(f) What does $E(\bar{y}) = \bar{Y}$ mean ?

(g) Give an example of a heterogeneous population.

(h) Define cost function in stratified sampling.

(i) Explain stratifying factor.

(j) Mention one disadvantage of stratified sampling.

ANSWERS

1. a- (iv) b- (iii) c- (i) d- (ii) e- (iii)
   f- (iv) g- (iii) h- (ii) i- (ii) j- (iv)

2. a- parameter b- inversely c- SRSWOR d- s^2/n e- self weighing
   f- True g- False h- False i- True j- False

***

APPENDIX - I

TABLE OF RANDOM NUMBERS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	1	2	5	4	2	8	5	8	7	3	5	8	4	0	2	4	3	6	8	4	8	4	8	5	2	6	1	7	5	4	8	8
2	5	4	4	3	4	9	1	1	0	9	2	2	7	1	3	4	4	7	9	8	1	3	1	1	8	7	0	1	2	2	1	0
3	3	2	6	2	2	3	2	2	4	1	1	2	9	8	7	7	4	7	7	6	4	5	1	2	1	7	4	6	2	5	9	3
4	7	8	0	9	0	2	9	7	8	9	5	6	2	1	5	8	7	7	8	0	0	7	5	3	1	2	3	2	7	1	8	1
5	6	8	6	2	4	1	9	4	3	5	9	6	5	0	7	2	4	4	7	3	3	0	9	9	0	7	2	9	4	9	5	0
6	9	1	7	9	3	8	1	4	9	1	5	3	2	1	2	7	6	7	4	5	9	6	4	6	8	1	0	5	3	1	3	3
7	5	3	1	7	0	9	8	6	0	6	3	3	6	4	8	0	4	8	3	4	8	7	1	0	8	8	2	9	8	5	7	2
8	0	1	2	6	4	7	7	7	8	0	3	4	9	2	1	7	2	1	2	8	2	2	3	2	5	0	3	9	8	6	3	7
9	2	3	7	2	7	7	7	4	9	4	4	6	7	1	7	8	8	4	0	3	3	9	7	1	0	8	9	9	5	2	7	4
10	0	3	5	7	5	2	7	6	3	9	9	9	0	2	6	1	9	2	5	5	5	7	8	0	5	7	2	8	0	0	3	2
11	7	8	5	5	9	7	0	7	5	2	5	9	4	2	6	3	9	8	7	8	4	9	1	8	0	9	8	7	9	1	1	8
12	2	5	1	0	4	2	5	4	1	5	4	3	0	2	2	4	0	1	1	2	6	5	2	3	8	6	6	7	4	7	0	7
13	6	6	3	9	1	9	1	3	3	1	2	0	9	1	4	9	6	1	4	5	5	8	9	5	0	7	2	6	3	8	8	3
14	6	7	6	9	1	4	3	5	9	1	0	7	4	7	6	2	9	9	0	2	3	7	6	4	7	3	8	8	2	7	2	9
15	4	5	2	7	8	0	0	0	8	6	4	8	3	3	6	6	7	9	4	5	4	8	4	7	4	3	1	7	9	6	3	6
16	5	6	9	9	9	8	8	3	2	4	5	6	0	8	9	3	4	1	3	2	6	6	6	8	0	7	9	9	6	1	3	7
17	4	1	6	0	1	4	4	5	2	8	8	7	0	7	2	4	1	2	9	4	8	9	8	8	1	5	2	7	1	4	6	7
18	4	5	0	6	2	4	7	4	3	5	9	0	5	3	0	8	7	6	4	0	7	1	2	8	1	0	2	3	2	4	1	8
19	4	6	4	5	0	6	1	3	9	8	4	6	4	4	5	8	5	6	6	6	7	6	7	1	1	1	8	4	2	3	2	8
20	6	6	8	6	3	5	4	4	9	8	2	8	9	1	8	7	0	5	0	6	6	4	7	3	5	3	5	6	8	9	4	0
21	6	5	0	3	0	3	2	9	7	8	9	9	8	2	1	1	0	8	5	2	8	0	6	6	5	7	0	6	1	9	4	0
22	1	6	5	8	4	2	8	8	1	8	5	6	5	3	1	9	3	5	1	2	8	9	8	1	7	4	6	8	3	8	3	6
23	1	3	1	9	1	2	0	4	3	3	4	4	8	8	8	6	3	9	4	6	6	7	7	7	1	7	0	0	0	3	2	3
24	5	0	3	0	4	0	2	7	2	0	7	7	9	8	1	2	8	6	4	5	3	2	9	0	7	1	9	1	6	1	5	2
25	7	8	6	6	2	0	2	9	5	1	5	6	2	0	0	3	2	9	4	0	0	2	3	7	7	6	7	0	2	8	5	2

APPENDIX - I
TABLE OF RANDOM NUMBERS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
26	1	9	8	3	8	9	9	2	1	0	1	7	7	2	6	3	7	6	9	9	4	1	5	1	8	1	3	2	7	2	7	1
27	9	9	4	4	0	8	4	5	7	4	6	8	3	9	3	6	8	0	0	2	0	8	5	7	5	7	8	4	4	4	8	0
28	0	3	3	0	9	9	1	3	4	9	9	0	7	7	9	0	6	9	3	2	0	8	7	1	1	9	5	8	9	8	8	1
29	9	9	0	3	7	9	1	4	4	1	3	8	6	8	2	6	0	2	3	0	1	3	3	7	7	4	1	3	8	8	4	0
30	1	6	1	4	7	8	6	2	9	5	0	0	4	1	0	9	1	0	3	7	2	9	7	8	6	0	7	5	0	9	7	1
31	2	0	9	6	1	1	6	4	1	7	8	8	6	2	5	7	0	6	3	2	0	6	9	3	2	2	6	3	5	2	9	0
32	0	5	1	1	0	2	2	9	5	9	5	1	6	8	0	8	1	4	0	9	7	6	2	3	4	0	6	1	4	6	9	7
33	0	5	2	4	4	0	5	6	9	1	4	0	6	3	7	1	2	0	9	9	8	2	9	0	3	6	1	1	6	5	0	1
34	7	3	8	1	7	3	8	6	6	5	6	8	1	5	6	8	4	1	6	0	0	4	2	9	3	4	8	8	3	7	4	1
35	3	3	1	1	3	7	3	3	7	8	8	2	6	9	8	5	7	8	7	4	7	2	6	4	4	5	8	7	9	5	9	1
36	6	8	7	4	2	5	3	4	7	4	8	5	9	5	9	6	9	0	8	6	2	7	0	1	4	9	6	7	1	5	8	8
37	8	9	8	7	3	1	2	1	3	6	2	8	0	3	7	2	1	0	5	9	6	3	3	9	8	9	7	3	4	2	1	8
38	9	2	0	5	2	3	5	8	0	3	9	3	3	1	9	6	1	6	1	2	8	3	9	7	4	3	9	0	0	1	8	7
39	9	7	4	9	7	9	8	3	7	0	7	6	9	7	9	1	1	5	3	0	8	1	2	7	4	4	7	4	1	8	9	5
40	2	1	8	3	2	1	0	9	2	8	7	4	5	7	3	3	1	5	6	7	7	7	6	4	4	9	3	9	9	9	1	9
41	6	9	2	6	3	0	8	5	2	0	7	9	3	3	3	0	4	4	3	2	9	5	2	4	0	3	2	7	9	6	4	0
42	3	3	7	3	3	5	6	7	0	3	7	1	5	9	3	2	3	9	2	3	7	2	5	0	8	5	7	8	5	8	6	9
43	9	7	7	1	5	5	4	2	4	7	1	5	5	5	2	7	3	7	6	3	3	1	6	7	3	6	7	9	4	3	9	9
44	9	3	5	3	5	5	7	6	5	4	7	4	0	1	9	0	7	2	7	4	6	9	9	3	3	9	2	0	7	2	7	2
45	5	7	6	1	6	3	0	1	3	5	5	8	6	2	0	5	3	0	1	2	6	1	9	5	8	4	6	1	2	0	4	6
46	6	8	5	0	8	1	2	2	4	4	5	5	2	9	4	0	9	9	4	5	9	6	8	8	3	5	8	8	8	3	1	1
47	6	6	1	6	6	7	6	0	4	9	3	8	0	0	6	6	0	3	9	1	8	8	9	8	4	7	5	3	7	4	0	2
48	2	6	3	3	4	2	5	5	8	7	5	5	8	4	3	4	4	3	3	4	8	9	9	2	1	2	6	0	0	5	4	7
49	6	8	3	7	4	8	9	8	9	5	2	7	3	5	2	6	6	5	3	6	7	7	0	3	9	9	4	1	8	4	5	4
50	4	5	2	3	7	7	7	2	3	8	8	8	7	2	4	3	1	3	3	0	7	1	1	5	4	8	9	7	6	6	1	8

CHAPTER-7

STATISTICAL SYSTEM IN INDIA

7.1 INTRODUCTION

Every government needs information regarding the social, economic, industrial, agricultural conditions prevailing in the country to frame suitable national policies. Such information collected and published by government agencies are called official statistics.

Kautilya's Arthashastra throws light on the system of collection of statistical information regarding land, area, agricultural production, population, taxes etc. that prevailed even 2000 years ago. During the Moghul period statistics were collected and used for administrative purposes. Evidences relating to statistical information during the reign of emperor Akbar have been recorded in 'Tuzke-Babri and Ain-i-Akbari.

Ryotwari system was introduced in some parts of India during the 18th century and the land revenue officers were appointed by East India company to collect land revenue. The revenue rate was based on agricultural production and fertility of soils. The procedure continued till the existence of zamindari system in India.

For the first time in India, in 1862, a statistical committee was formed to look to the compilation and collection of statistics relating to trade, finance, education, agriculture etc. The first Gazetteer containing economic statistics for provinces was prepared in 1866. In 1868, the Statistical Abstract for British India was first published from London. A partial census in India took place in 1872. In 1875, a Department of Agriculture and Commerce was set up in Uttar Pradesh for improving the agricultural statistics in the country. In 1881, Agricultural Departments were opened in various provinces and the centre as well, on the recommendation of the Indian Famine Commission for collecting valuable information relating to various agricultural problems. The first complete population census, covering the whole of India, was also conducted in 1881 systematically. Since then, decennial census has been a regular feature. The Imperial Gazetteer of India which contained economic statistics of different parts of the country was also published for the first time during 1881. The statistics of foreign trade, prices and industries, on the other hand, were collected and published by the Department of Finance and Commerce. This led to the publication of "Agricultural Statistics of British India" in 1886.

In order to collect and compile relevant agricultural data adequately, a separate. Statistical Bureau was set up at the Centre in 1885. But, in 1884, the first Crop Forecast of wheat production was made. In 1905, the Directorate of Commercial Intelligence and Statistics was established and the Statistical Bureau was merged with it. In 1906, the first issue of the Indian Trade Journal was published by this office.

The 'Economic Enquiry Committee' was constituted in 1925, under the chairmanship of Sir M. Visweswaraya, to enquire into the adequacy of the statistical data available and to recommend for their improvement. An important development that occured in 1933 was the establishment of Statistical Research Bureau for the purpose of analysis and interpretation of economic statistics. In 1934, the famous Bowley-Robertson Committee recommended and gave details of the possibility of an Economic Census in India. On the basis of the recommendation, the Government of India decided to set up the Central Statistical Organisation which could not be implimented at that time due to financial implications and practical difficulties. With the out break of the First World War in 1939, the need for statistical information was felt more and more. The Department of Industries and Civil Supplies, meant for controlling the civil supplies, was entrusted with the job of collection of the statistical information relating to essential commodities. To ensure completeness of data on industries, an act known as 'Industrial Statistics Act' was passed in 1942. In 1946, the first census of manufacturers was conducted. Thus, it is evident from the above facts that official statistics in India during pre-independence period was in a very poor state. The chief reason was the lack of interest of the British Government in India. Whatever interest they were taking was confined to their administrative convenience.

It was only after the Independence, steps were taken towards the economic development of the country through implimentation of successive five-year plans. In 1947, the Economic Adviser's office started publishing the general purpose wholesale price index numbers. In 1949, Prof. P. C. Mahalanobis was appointed as the Honorary Statistical Advisor to the government and on his recommendation a National Income Committee (NIC) was set up in the same year to estimate India's National Income year wise. In 1950, the National Sample Surveys (NSS) were conducted for the first time. A Central Statistical Unit was set up in the Cabinet Secretariat in 1949 and subsequently, in 1951 this unit became the Central Statistical Organisation (CSO).

In India, we have at present, a broadly decentralised statistical system in which the CSO, with its headquarters in New Delhi, acts as the apex or advisory and co-ordinating body.

Collection of Statistics is divided between the Central Government and the State Governments on a subject wise basis. Under a Federal Constitution, the CSO is responsible for the co-ordination at the national level for all the activities of the states and central statistical agencies under different Ministries, while in the state level, the State Statistical Bureaus act as the co-ordinating agency.

Constitution of India, under article 246, clearly specifies three lists viz. the Union List, the State List and the Concurrent List.

Under the Union List, Foreign Trade, Currency and Foreign Exchange, Banking, Railways, Post and Telegraphs, Defence, Population Census, Customs and Excise Duties and Income Taxes etc. are under the direct control of Union. The Government of India bears full responsibility and cost of collection of data on these items.

Under the State List, subjects like Public Health, Agriculture, Live stock, Irrigation, Forests, Fisheries, Education etc, are included. The State Governments bear the responsibility of data collection. for these items. But there are some subjects on the Concurrent-List like Industry, Trade Unions, Labour Disputes, Relief and Rehabilitation, Price-Control; in respect of which the central and the state governments operate simultaneously to meet their respective requirements of data. Where the States (and the Union Territories) have the primary responsibility for data-collection, the Central Government acts (Through the Central Statistical Organisazation, CSO) as the co-ordinating agency for the compilation and publication of data on all India basis.

7.2 STATISTICAL ORGANISATIONS UNDER THE UNION GOVERNMENT

For a systematic organization and collection of statistical information, Government of India established different departments at the Centre and State levels. At present, so far as the Centre is concerned, each Ministry has a statistical unit for collection and compilation of statistics relating to the Ministry. Some of the important organisations are:

1. Directorate of Economics and Statistics (DES)
2. Indian Council of Agricultural Research (ICAR)
3. Office of the Census Commissioner and Registrar General.
4. Labour Bureau.
5. Department of Commercial Intelligence and Statistics (DCIS)
6. Directorate of National Sample Survey (NSS)
7. National Income Unit. (NIU)
8. Research Section of Reserve Bank of India (RBI)
9. Central Statistical Organization (CSO)
10. Directorate of Industrial Statistics (DIS)

We describe some of the important statistical organisations in the following :

7.2.1 Central Statistical Organisation (CSO) :

The CSO was set up by the government of India in 1951 as a part of the Cabinet-Secretariat with the objective of creating co-ordination of large variety of statistical information, collected at the Centre and State level. The functions of the CSO expanded by the transfer of the National Income Unit from Ministry of Finance in 1954 and the Directorate of Industrial Statistics from the Ministry of Commerce and Industry in 1957. The status of the CSO was raised to that of a Department in 1961 under Cabinet Secretariat.

Functions :

The functions of the CSO are :

1. It co-ordinates the statistical activities at the Centre and the State.
2. It plays advisory role in statistical matters and provides national statistics to the United Nations and its specialized agencies and other international bodies.
3. It brings out publications presenting all - India statistics on all principal aspects of national life.
4. It attends to the statistical work relating to the five-year plans through its separate unit in collaboration with the Planning Commission and organises and conducts training courses in statistics for both, those who are in Government service and outsiders.
5. It compiles and publishes National Income Statistics.

6. Through its Industrial Statistics Wing, it conducts the Annual Survey of Industries and publishes the results.
7. It takes steps to set and improve the standards regarding concepts, definitions, classification and methodology of data collection.
8. It takes the responsibility of construction of consumer's price indices.
9. It presents economic and social statistics graphically.

The CSO also co-ordinates with National Sample Survey Organization. It provides data for publication in the

(i) U. N. Monthly Bulletin of Statistics.
(ii) U. N. Quarterly Bulletin on Commodity Trade Statistics.
(iii) U. N. Demographic Year Book.
(iv) Economic Council for Asia and Far East (ECAFE) Quarterly Bulletin and Annual Surveys.

Besides, the other functions of the CSO are to

(i) co-ordinate the conduct of annual survey of industries in all the States, and process the data for publication.
(ii) give monthly account of the production of selected industries through a monthly bulletin.
(iii) compute and publish the monthly index of industrial production obtained for selected industries and
(iv) function in many other ways.

The CSO renders advice to various ministries in the Central Government. It examines the bulletins of the State level. It lays down the definitions and concepts of data collection in conformity with international standards viz. International Standard Industrial Classification (ISIC); International Standard Classification of Occupation (ISCO) and Standard Occupational Classification (SOC) for India.

Publications :

Statistical information are collected from a vast field of economic activity like motor vehicles, civil aviation, foreign trade, inland trade, taxation and revenue, agriculture, forest, fishery, livestock etc. and are compiled and published by the CSO. Some important publications of the CSO are :

1. **Monthly Abstract of Statistics :**

It represents data about different facts of Indian economy viz. labour and employment, fuel and power, minerals, industrial production, transport, foreign trade, banking and currency, prices, consumption and stocks, postal traffic etc.

2. **Statistical Abstract - India (Annual) :**

It gives an account of area, population, climate, agriculture, mining, banks, motor vehicles, balance of payments etc. Large number of tables are presented relating to social, economic and natural factors.

3. **Annual Survey of Industries :**

The National Sample Survey Organisation carries out annually a survey of industries known as Annual Survey of Industries (ASI). Industry wise estimates of employment, output, input and capital are displayed separately for the census and sample sectors.

4. **National Accounts Statistics (NAS) (Annual) :**

This annual publication is also known as white paper. It incorporates the estimates of capital formation, savings, private consumption, expenditure and the disaggregated tables.

5. **Monthly Statistics of the Production of Selected Industries :**

It gives the monthly production statistics. Two month's statistics are published in a combined issue covering a large number of items. It also contains the index of industrial production & mill stock position.

6. **Statistical Pocket Book (annual)**
7. **Sample Survey of Current Interest in India (annual)**
8. **Estimates of National Products (revised series) (annual)**
9. **Basic Statistics relating to the Indian Economy (annual)**
10. **Retail Price Bulletin (monthly)**
11. **Statistical News Letter (quarterly)**
12. **Statistical System in India (adhoc)**
13. **Official Statistics (adhoc)**
14. **Estimates of Savings in India (adhoc)**
15. **Estimation of Capital Formation in India (adhoc)**

Divisions :

The CSO performs all its functions cited above through its various divisions given below.

1. Industry and Trade Divisions.
2. Industrial Statistics Wing.
3. Training Division.
4. Family living Survey Division.
5. Population Division.
6. National Income Division.
7. National Sample Survey Division.
8. Manpower Research Division.
9. Planning and State Statistics Division.
10. Statistical Intelligence Division.
11. Analytical Division.
12. Price and Cost of Living Statistics Division.

7.2.2 National Sample Survey Organisation (NSSO)

The National Sample Survey Directorate was initially set up in the year 1950 in the Ministry of Finance at the initiation of Prof. P. C. Mahalanobis. The object was to compile sample data on a continuing basis, needed on all aspects of national economy, as required by the National Income Committee (NIC), the Planning Commission and various Ministry of the Government of India. It was brought under the control of the Department of Statistics, Cabinet Secretariat in 1957. In 1971 January, National Sample Survey Organisation (NSSO) was created as a part of the Department of Statistics with a view to bring about better programming and effective co-ordination. The Directorate of the NSS is now a part of the NSSO and is renamed as Field Operations Division (FOD).

Functions :

The main functions of NSSO are to

(i) conduct socio-economic surveys with an all India coverage by the method of random sampling for collecting data on socio-economic conditions of the people, prices and wages, production in small scale household enterprises, consumption and agriculture.

(ii) conduct annual surveys in the organised industrial sectors.

(iii) provide technical guidance for the conduct of crop estimation surveys and crop cutting experiments.

(iv) provide statistical data for national income and planning.

(v) train personnel and provide guidance to the states for the conduct of surveys.

(vi) evolve statistical techniques to bear on the analysis of information, the solution of administrative problems and the estimation of future trends.

(vii) provide and analyse information which are useful to the research workers.

(viii) assist in keeping the public informed about the new developments in the economic and the social fields.

The activities of NSSO are controlled by a Governing Council with regard to survey designs, field operations, data processing, economic analysis and publication of NSS data. In case of the socio-economic surveys, decision on subjects to be covered is taken on a round wise basis, keeping in view the request for data from the Central and State Governments. It conducts surveys on demography, health and family planning, debt and investment, capital formation, land holdings, livestock enterprises (quinquennially), employment, rural labour, consumer expenditure and self employment in non-agricultural sectors. NSSO also conducts surveys on special demands.

Divisions:

The present structure of NSSO consists of four functional divisions, with a chief executive officer at the apex. The four divisions are :

(a) Survey Design and Research Division

(b) Field Operations Division (FOD)

(c) Data Processing Division

(d) Economic Analysis Division

The Chief Executive Officer, as the member Secretary of the Governing Council obtains the approval of the Council for the programme to be undertaken and directs the appropriate division to implement it. The tabulation and analysis of data are attended by the CSO and, partly, by the Labour Bureau and National Building Organisation (NBO). So-far as the agriculture is concerned, the Field Operations Division (FOD) is responsible for all activities.

NSSO conducts training programmes for its personnel and accordingly, 41 regions of FOD have been grouped into five zones, with headquarters at Bangalore, Nagpur, Jaipur, Allahabad and Kolkata. Each zone has been placed under the charge of a Deputy Director who is responsible for organising training courses for his zonal staff.

The important publications of the NSSO are the Reports on the Various Rounds of the NSS and The Quarterly Bulletin "Sarvekshana".

7.2.3 Office of the Registrar General

In 1949 the Government of India decided and set up in the Ministry of Home Affairs, an office, named the Office of the Registrar General of India (RGI). The purpose of this office was to pay continuing attention to the work of the decennial censuses which had till that time been entrusted to a Census Commissioner appointed on an adhoc basis. With the creation of the permanent post of Registrar General, the Registrar General started functioning as ex-officio Census Commissioner.

The main publications of the RGI are :

(i) The Census of India Reports,

(ii) Vital Statistics of India (annual)

(iii) Sample Registration Bulletin (quarterly)

7.2.4 Office of the Directorate General of Commercial Intelligence and Statistics (DGCIS)

This organisation was set up in Calcutta (Kolkata) in 1895. The DGCIS is responsible only for commercial intelligence and trade statistics. The licensing statistics and the balance of trade statistics relating to the country's external trade are, however, handled by the Chief Controller of Imports and Exports (CCIE) and the Reserve Bank of India (RBI) respectively.

The main publications of the DGCIS are :

(i) Monthly Statistics of Foreign Trade of India (in two volumes)

(ii) Indian Trade Journal (weekly)

7.2.5 Directorate of Economics and Statistics (DES)

This Directorate was set up in 1947 in pursuance of the decision of the Union Government to centralise all services relating to agricultural economics and statistics. Now attached to the Ministry of Agriculture and Irrigation, the DES is the central co-ordinating agency which is responsible for the collection, compilation and publication of agricultural statistics at the all-India level. The data cover, besides agriculture (i.e. land utilisation, area under crops and crop production), !ivestock, forests and fisheries.

The important publications of the DES are :

(i) Indian Agricultural Statistics (annual)

(ii) Estimates of Area & Production of Principal Crops in India (annual)

(iii) Agricultural situation in India (monthly)

(iv) Indian Forest Statistics (annual)

(v) Bulletin on Food Statistics (annual) and

(vi) Indian Livestock Census (quinquennial)

Besides, DES also publishes the following :

(1) Indian Land Revenue Statistics.

(2) Indian Agriculture in Brief.

7.2.6 Labour Bureau, Simla

This office was set up in 1946 in the Ministry of Labour and Rehabilitation. The Bureau has the following main functions.

(i) It collects, compiles and publishes statistics of employment in respect of factories, mines, plantations, shops, commercial establishments, etc, on all India basis.

(ii) It constructs consumer price index numbers.

(iii) It collects data necessary for the formulation of government policies.

(iv) It brings out pamphlets on different aspects of labour legislation.

The following are the important publications of the Bureau.

(1) Indian Labour Statistics (annual)

(2) Indian Labour Year Book (annual)

(3) Indian Labour Journal (monthly)

(4) Employment Review (annual)

(5) Agricultural Wages in India (annual)

(6) Working of the Trade Unions Act (annual)

(7) Statistics of Factories (annual)

7.2.7 Indian Statistical Institute (ISI)

The ISI was established in 1932 at Calcutta (now Kolkata) as the pioneering institute for the development of the statistical system in India. Its status was conferred to the status of an institution of national importance through an Act of Parliament in 1959. It helps in the research, training and application of statistical methods to a wide range of problems such as study of rainfalls, survey of agricultural crops, floods and socio-economic enquiries. It holds examinations and awards certificates, diplomas and degrees of proficiency in statistics as any other university. It also takes up technical works for the NSS such as designing of the survey, preparation of the schedules, tabulation of data and report writing. This institute also conducts the international statistical education centre at Kolkata, a nine months training programme, for students from Asian countries, in collaboration with the UNESCO and the International Statistics Institute. It is functioning as a focal centre for professional training and research and as the national statistical and computational laboratory in India. The institute is also bringing out "The Sankhya", a statistical journal of international repute.

7.2.8 Research Department of Reserve Bank of India (Mumbai)

The research department of the Reserve Bank of India (RBI) was set up in 1945 and discharges a number of functions through its four different divisions as under :

(i) Monetary Research Division :

Through this division, it prepares the annual report of the Bank and the report on Currency and Finance. This division is concerned with the processing of statistics of Banking Stock Exchange, Bullian Markets and Public Finance.

(ii) Balance of payments Division :

Through this division, it undertakes a continuous assessment of the country's external and internal economic situation including a census of foreign liabilities and assets.

(iii) Statistics Division :

Through this division, it issues a monthly bulletin (RBI bulletin) bringing out the salient facts that emerge out of the country's internal and external economic situation.

(iv) Rural Economic Division :

Through this division, it undertakes surveys on agricultural indebtedness, rural finance etc.

7.2.9 Indian Council of Agricultural Research (ICAR)

The statistical division of ICAR was founded in the year 1931. At present this division is being renamed as the Institute of Agricultural Research Statistics (IARS). Its main functions are as under :

(i) Use of the modern experimental designs in biological research.

(ii) Introduction of the method of random sampling for the estimation of yield of crops.

(iii) Introduction of suitable designs for experimentation in the field of agriculture.

(iv) Conducting research and imparting training and awarding certificates and diplomas in agriculture and animal husbandry statistics.

7.3 STATISTICAL OFFICES IN THE STATES

As a result of the Gregory Committee in 1946, Statistical Bureaus were established in most of the States of India. In Odisha, the Bureau of Economics and Statistics was set up. It is now called the Directorate of Economics and Ststistics. These Directorates carry out the following main functions.

(1) Co-ordination of statistics collected by different departments of the State Government.

(2) Publication of abstracts assembling all essential statistical series.

(3) Maintenance of liasion between the statistical units in the state departments on the one hand and the CSO and other statistical offices at the centre on the other.

(4) Organisation of special inquiries and surveys.

(5) Compilation of economic indicators and income statistics for the state, and

(6) Undertaking statistical work relating to planning.

Practically, all the Directorates are now participating in the Socio-economic Surveys conducted by the NSSO.

The principal publications of a State Bureau are (i) The Statistical Abstract (annual) and (ii) The Statistical Bulletin (monthly or quarterly).

At the district and regional levels, information on a variety of statistical data are available in the form of documents, reports, bulletins, hand books etc. The published data are to be found in the District Gazetteers, District Census Hand Books, District Statistical Abstracts, State plan documents, publicity booklets of Director of Public Relations, State Livestock Census Reports, Lead Bank Survey Reports, Price Bulletins, the Municipal Year Book etc. The statistics collected by the State Planning Boards for formulation of the district plans are too voluminous to be published.

7.4 AGRICULTURAL STATISTICS

As a matter of fact, all statistics which have an impact on agricultural economy may be regarded as agricultural statistics. At the all India level, the Directorate of Economics and Statistics under the Ministry of Agriculture and Irrigation (DES-Ag) is the central co-ordinating agency responsible for the collection, compilation and publication of agricultural statistics. The data are collected mainly by State Governments and supplied to the DES for compilation and publication.

The data on agriculture can be classified under four broad heads.

(i) Land Utilisation

(ii) Area & Yield including forecast of yield.

(iii) Agricultural wages and prices.

(iv) Statistics regarding livestock, poultry, forestry and fisheries etc.

7.4.1. Land Utilisation Statistics :

These are statistics giving information about the areas of land put to different uses, areas irrigated and crops irrigated and unirrigated areas under different crops. Currently, land use statistics are available for about 92% of the tota' ȝeographical area of the country.

Till 1949-50, land use statistics used to be presented according to a five - fold classification. A nine-fold classification replaced the old classification in 1950-51. The correspondence between the old and the new classes and the descriptions of the new classes are given below :

CLASSIFICATION OF LAND USE STATISTICS

Old class	**New class**	**Description**
1. Forests	1. Forests	The class includes all actually forested area or lands classed or administered as forests under any legal provision, whether state- owned or private.
2. Area not available for cultivation	2. Land put to non-agricultural use.	Land occupied by buildings, factories, roads, play grounds, railways or land under water or land put to uses other than agricultural.
	3. Barren and unculturable land.	Lands like mountains and deserts and land which can not be brought under cultivation except at a high cost.
3. Other uncultivated land excluding current fallows.	4. Permanent pastures and other grazing lands	All grazing lands, whether they are permanent-pastures and meadows or not.
	5. Land under miscellaneous tree crops and groves	All cultivable land which is not included under net area sown, but is put to some agricultural use.
	6. Culturable waste	All lands available for cultivation, but not taken up for cultivation or, even if taken up, abandoned after a few years for some reason or the other.
4. Current fallows	7. Current fallows	Cropped areas which are kept fallow during the current year.
	8. Other fallow lands	All lands which were taken up for cultivation but have been temporarily out of cultivation for a period of not less than one year and not more than 5 years.
5. Net-area sown	9. Net area sown	Net area sown with crops and orchards, the areas sown more than once in the same year, but counted once only.

From the stand-point of collection of area statistics, the States may be divided into three groups. In the first group are the former temporarily-settled States where, the village revenue agency maintains land utilisation statistics as a part of land records. These are collected by the patwaris on the basis of complete field-to-field enumeration and are fairly reliable. The second group comprises the former permanently-settled States of West Bengal, Odisha and Kerala, where no village revenue agencies exist. These States have adopted the sample survey method for obtaining land use statistics. The third group consists of areas which are neither cadastrally surveyed nor have any revenue agency. For these areas, the statistics reported are in the nature of eye estimates of revenue officers.

Statistics of area irrigated are also collected as a part of land use statistics. This area is classified both according to source of irrigation (canals - Government and private, tanks, wells and other sources) and according to crop irrigated. In case two crops are irrigated from the same source on the same land in the same year, the irrigated area is classified by 'source represents the net irrigated area', while the area irrigated is classified by 'crop represents the gross irrigated area'.

Publications :

Each state publishes its own land use statistics in Season and Crop Reports. But the DES compiles the all - India figures and publishes in two-volumes. These are :

Indian Agricultural Statistics (annual) - Vol. I and Vol.-II. In Vol I the state -wise and in Vol II the district-wise data are published. All India summary tables are given in both the volumes. Vol.-II in addition, gives an introductory note concerning the salient-features of rainfall, land use, irrigation and cropping pattern of the year.

Land utilisation Statistics are also published in

(i) Agricultural Situation in India (monthly)

(ii) Abstract of Agricultural Statistics (DES, annual)

(iii) Statistical Abstract of the Indian Union (CSO, annual)

(iv) Indian Land Revenue Statistics.

(v) Indian Agriculture in Brief.

7.4.2 Area Statistics :

Each state has a Directorate of Economics and Statistics which collects statistics regarding area, crop production, poultry and forests of the State. Total area statistics are obtained from two sources, (i) The Surveyor General of India and (ii) The village records maintained by the Revenue Department. The area is further classified as forest, land put to non-agricultural use, barren and unculturable land, permanent pastures and other grazing lands, culturable waste, current fallows, other fallow land, new area sown (irrigated and unirrigated). The irrigated area is found out by combining the area under irrigation by canals, tanks and wells.

The area under crops is obtained by two sources.

(i) Official Series, based on village records. The official series relate to the statistics of the area under land utilisation for different crops.

(ii) NSS Series, based on sample surveys.

The Directorate of NSS (Now NSSO) collects data during the regular rounds of surveys on area under different crops. The area figures provided by these two sources differ widely zone wise because of the difference in method of the coverage of crops, difference in the field work, difference in the classification of area under grain crop and fodder crop, the allocation of area under mixed crop and due to sampling error in NSSO estimates.

Area sown with a crop is taken to mean the area actually sown, no matter, whether the crop reaches maturity or not, except in cases where the land is devoted to another crop following the failure of the first crop. In the latter cases the area is shown under the next crop and is excluded from the first crop. If no other crop is sown, then the area is shown under the old crop.

7.4.3. Crop production & Yield Statistics :

Agricultural production includes the production of food and non-food crops. (excluding forests, livestock and fisheries)

(a) Food crops - (i) foodgrains (cereals and pulses) (ii) sugarcane (iii) condiments and spices, (iv) fruits and vegetables, and (v) other food crops.

(b) Non-food crops - (i) oilseeds, (ii) fibres, (iii) dyes and tanning material, (iv) drugs, narcotics and plantation crops, (v) fodder crops, (vi) green manure crops, (vii) guar and oats, and (viii) other non-food crops.

Crop output is estimated by multiplying the area under a crop by the average expected yield per hectare, in the season. The yield statistics are collected by official machinery and NSSO. As official machinery, the work of the survey is under taken by the Directorate of Economics and Statistics.

For the estimation of yield (formerly known as crop forecast), the various crops are divided into two groups viz. (i) forecast crops and (ii) non-forecast and plantation crops. In India, two methods are adopted to collect yield statistics of various crops :

(a) Traditional Method

(b) Random Sampling Method

7.4.3.(a) Traditional Method : (Annawari Method)

Estimates of yield of 38 crops, including foodgrains, oil seeds, fibres and crops like potato, sugar cane, tabacco etc. are those for which regular all-India estimates of area and production are issued. The periodical estimates of area and production are initially prepared by the concerned State agencies but are complied by the DES and issued on pre-assigned dates. For each of these crops, usually two to three estimates are issued; while for cotton five estimates and for castor seed only one estimate. The first forecast is based on the general impression and usually issued one month after the sowing of the crop. The second comes about two months later and includes the area of late sowings and is based on the general condition of the crop. The final forecast, however, attempts to provide firm estimates of the total area sown and the total production. These are revised about one year and about two years later in the light of returns received from defaulting states.

Computation of yield Estimate :

The estimation of the yield is done by using the formula

Total yield = Area under the crop (in hectares) x yield per hectares

= Area under the crop (in hectares) x Normal yield x Condition factor

Some authors prefer to writ ,

$$\text{Condition factor} = \frac{\text{Condition factor judged for the crop in the year (in annas)}}{\text{Normal yield condition (in annas)}}$$

Thus the traditional method of estimating the yield per hectare, called the Annawari method, is based on the notions of the 'normal yield' and the 'condition factor'. The 'normal yield' refers to a district and is defined as the average yield on an average soil in a year of average character. The average yield is fixed on the basis of crop-cutting experiments in the field selected by revenue officers as bearing average. The condition of the crop is judged through eye estimation by Patwaris or Lekhpals or by any other authorised officer of the Agriculture or Revenue Department. This is known as the "condition factor" or the Annawari estimate. The 'condition factor' refers to a village and is taken to reflect to what extent the village y'eld per hectare during the given year is likely to differ from the normal yield. The factor is expressed as so many annas per rupee, the rupee representing the normal yield.

Example : To calculate the estimated average yield of paddy crop in Odisha, we have used the following imaginary information:

Normal yield condition is 11 annas

Normal yield of the paddy crop is 15 quintals per hectare

Condition factor of the crop in the year is 8 annas.

$\therefore$ Average estimated yield per hectare = 15x8/11 = 10.90 quintal = 11quintals.

This system suffers from two main lacunae. Firstly, the condition factor depends totally on the vizual judgement of lekhapals or patwaris, or the officer concerned and thus is very subjective. Secondly, the total annas representing the normal condition of a crop is not uniform in all the States. Therefore, the traditional method is now being abandoned.

7.4.3.(b) Random Sampling Method :

This method was recommended by the Board of Agriculture as early as 1919 but was abandoned due to the huge financial burden on the States. In 1923-25 Sir J. A. Hubback adopted the system of random sampling method to conduct a number of surveys on yields of paddy crops in Bihar and Odisha. The Indian Council of Agricultural Research introduced the random sampling method again in 1942 for cotton crops. This method was later extended to other crops in all States, except West Bengal and Odisha, where the job was entrusted to Indian Statistical Institute, Calcutta.

Currently, for most food crops and some cash crops (crops grown for selling, rather than for use by the grower), the estimation of yield rate is done with the help of crop-cutting experiments. The estimate is built up by actually harvesting, threshing, winnowing and weighing the crop growing in small areas (called cuts) selected among the fields.

In random sampling method, villages are randomly selected in tehsils of a State and within each village few fields are selected randomly. This is called stratified multistage random sampling method of selection. Here, the tehsils (each containing 100 to 300 villages) are strata, each village is a primary unit, a field growing the particular crop is the secondary unit and a cut within the field is the ultimate sampling unit.

For each crop, generally 2 to 10 villages are chosen at random from each stratum; in each village 2 fields growing the crop are selected; and in each field a cut of a prescribed size is marked for conducting the crop-cutting experiment. The size of the cut varies from 1/500th of a hectare (10m.x2m.) to 1/50th of a hectare (20m.x10m.) in the case of cotton. But the commonest cut - size is 1/200th (10m.x5m.) of a hectare.

The methods used in Kerala, Odisha and West Bengal are slightly different, where the job is entrusted to Indian Statistical Institute, Calcutta.

For non-forecast and plantation crops, the available estimates are adhoc estimates. The estimates of area and production of tea, coffee and rubber used to be based on special returns received by the DES from the State Governments.

Now crop-cutting experiments are conducted by the DES and Statistical Section of the Board of Revenue, through its field staff and separately by National Sample Survey Organisation (NSSO). The NSSO conducts crop-cutting experiments on major cereal crops during the course of their regular survey rounds.

Publications :

The two most important DES publications on area and yield of crops are the following :

(1) Estimates of Area and Production of Principal Crops in India (annual)

(2) Agricultural situation in India (monthly)

Other publications on agriculture are :

(1) Indian Agriculture in Brief (DES, annual)

(2) Bulletin on Food Statistics (DES, annual)

(3) The weekly Bulletin of Agricultural Prices

(4) Tea Statistics (annual), published by the Tea Board

(5) Coffee Statistics (annual), published by the Coffee Board

(6) Indian Rubber Statistics (annual) and Rubber Statistical News (monthly), both published by the Rubber Board.

Advantages of Random Sampling Method :

Random sampling method is scientific and reliable because of the following reasons.

1. The estimates obtained by stratified multistage random sampling method are free from personal bias, prejudice and whim of the investigator as the selection of the villages and the plots are done at random.

2. The estimates are based on the modern statistical techniques of sampling and so are expected to be true.

3. The standard error of the estimate can be measured.

7.5. POPULATION STATISTICS :

Population Statistics are a part of the science of demography. These are a set of figures relating to the number of people, their birth place, nationality, age, sex, marital status, and economic characteristics that a country possesses. It can be studied under three different phases. These are :

1. Population Census :

It means an official count of all the people either physically present or regularly residing in a given region at a given point of time. It includes, within its scope, the collection of information on various aspects of the people counted such as:- age, sex, race, religion, marital status, educational level, income etc.

2. Vital Statistics :

It is a branch of statistics, which deals with the registration of facts concerning birth, marriage, divorce, sickness and death and are collected under the State direction.

3. Demographic Surveys :

Ad hoc surveys, conducted by official agencies, of a particular region to gain the required information about population growth are called demographic surveys.

7.5.1. Methods of Conducting Population Census :

Population census in India dates back to 1872 when a partial census was held. The first complete census of the whole country was taken in 1881. Before the independence, Census in India was not conducted under a permanent legislation. But after the independence, Government of India passed an act, known as the Indian Census Act, 1948 which empowers the Central Government to take a census, appoint necessary census staff, delegate authority and the status of public servants on census officers, fix duties of census officers, assume the power to call upon persons to give assistance and the authority to ask questions and fix liability upon the citizens to furnish correct answers. The Act also authorises the Central Government to make rules and suspend other laws in respect of census by municipalities.

As a consequence of this Act, the census officers appointed legally are authorised to ask any question relevant with the census to any citizen and the citizen is bound to furnish the correct answer. The Act also authorises the census enumerators to enter a house to interrogate the people for the required information.

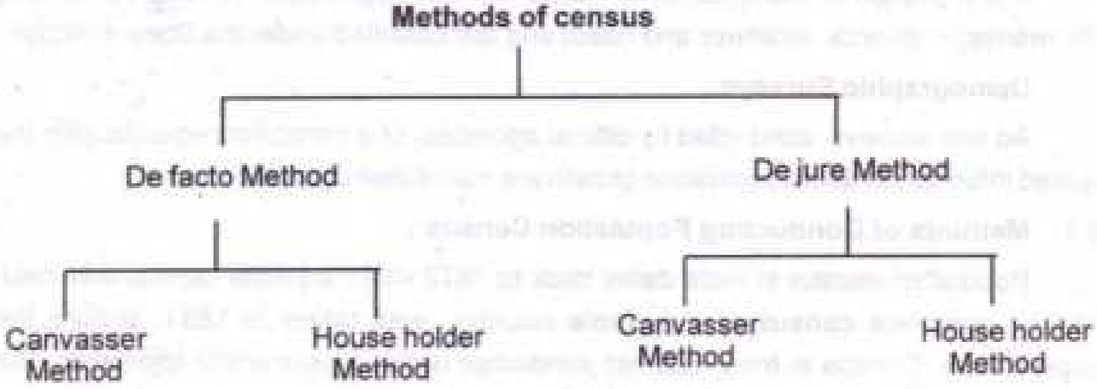
There are two distinct methods of census-taking : the canvasser method and the householder method. Under the first method, an enumerator approaches every household allotted to him and records the answers on the schedules himself after ascertaining the particulars from the head of the household, or any other knowledgeable member of the household. Under the second, the enumerator distributes the schedules to the households in his jurisdiction and the head of each household is expected to fill the answers in respect of all the household members and the enumerator collects back the answered schedules after the census date.

In fact, there are two methods of census. These are :

(1) De facto method or date system

(2) De jure method or period system.

Further, each of the above two methods can be adopted in two ways as follows :



De facto method : In this method, persons are counted, wherever they are present on the census night. Counting takes place on a single moonlit night of a particular day simultaneously all over the country. This is why it is called the date system or one night system. Enumeration of de facto population is difficult unless the movement of the population is restricted on the census night. Again, as the entire operation is completed in a single moonlit night, it is operationally very difficult. Till 1931, the De facto canvasser method used to be adopted in India for population census.

This method of census is simple because it does not require elaborate instructions. Further, since most of the countries use this method of census, international comparison is possible.

De jure Method :

In this method, people are counted on the basis of their habitual residence. Any person found temporarily at a place is excluded from enumeration and is counted by an enumerator at his usual place of residence, even if he is absent temporarily. In this method, census is not conducted on a fixed date but is conducted during a particular period. That is why, it is also known as the period system. The De jure canvasser method provides geographical and regional distribution of population accurately in comparison with the De facto method.

It is less complicated and is supposed to be complete. But this method requires detailed and elaborate concept of the definitions of the terms used. Further, in this method there is chances of counting a person more than once.

Because of the simplicity of this method the De jure canvasser method has been adopted in India since 1941 and the period of the census has been extended to three weeks.

Slip system :

Instead of filling the schedules, the slip system of census was introduced for the firs: time in India during the census of 1941. In this system, one slip was assigned to each individual on which the entire information about the person concerned was noted. Answers to various questions were recorded in the form of symbol codes on the slip. This system of recording answers saves a lot of labour and time of the enumerator and eases the processing of information with the help of mechanical aids.

7.5.2 Census Organisation :

The first census of free India, was conducted in 1951. Prior to 1951, there was no permanent census organisation. A census commissioner used to be appointed about - 18 months before the census date and the whole census organisation used to be disbanded soon after the census operation was over. With the creation of the post of Registrar - General of India (RGI), in 1949, the RGI, as ex-officio Census Commissioner, became the permanent authority to conduct all future censuses. Before every census, a Director of census operations is appointed in each State (or Union Territory), who is kept in charge of overall supervison of the census work in the area under his jurisdiction. Below the Director, there are District Census Officers. Again, each district is divided into a number of Charges, each under a Charge Superintendent. Each charge is divided into a number of Circles under a Circle Supervisor. Finally, each Circle is divided into a number of Blocks, and an Enumeretor is appointed to collect data for the whole Block in his charge. Normally, patwaris and school teachers are appointed as Enumerators in rural areas and school teachers and local officials are appointed as Enumerators in urban areas.

Conduct of Census :

It was for the first time in 1951, that a distinction between a House and a Household was made. A 'House' was defined as a dwelling place with a separate main entrance while a 'Household' was defined on the basis of 'Chulha' i.e. persons or group of persons living together and dining from a common kitchen were regarded as the member of the same household. This helped in estimating the family size. 1951 population census was conducted on the basis of household.

In the first stage, the census operation is conducted some months before the actual enumeration. All houses and households living in them are listed in the houselist schedules, along with related information. Using the data on household size, Blocks are formed, each comprising a population of about 750 in rural areas and about 600 in urban areas. At the time of the enumeration, the house list pertaining to each Block is updated.

During the enumeration period, enumerators visit all households in their respective Blocks and fill in the 'Household Schedules' as well as the individual slips and during the next few days, they go for a revisional round. In this round they visit all the households making necessary corrections for births and deaths, if any, occuring upto the reference date. They also include persons arriving in the households in the meantime who have not been enumerated else where. The houseless people (e.g. the pavement -dwellers and nomads) are enumerated on the last night of the census period (i.e., on the night preceding the census date) at the place where they are found.

Certain information, such as name, relationship to head of the household, sex, age, marital status, mother tongue, religion, occupation, literacy and educational attainment etc, are usually included in the census schedule. To acquire more information about the economic characteristics of the population, the changes were made in the slip of 1941 for the 1951 census.

Census 1961 :

In this census, De jure Canvassing Census Method was adopted. Each individual was contacted at his place of residence. Each house was identified by a location code. In census 1961, three categories were defined instead of two categories of houses, namely (i) Building (ii) Housing (iii) Household. The word building referred to an entire structure raised on the ground whereas, a census house referred to either a whole or a part of a building. Hence, a building may have one or more census houses. Information on the purpose for which the house was used was recorded. (e.g. dwelling, shop, shop cum dwelling, business, factory workshop, school or other institution, etc. or lying vacant). Also the predominant material of the roof and the wall were recorded. The questionnaire for 1961 census consisted of two parts namely (i) household schedule (part -I) and (ii) the individual census slip (part - II). Besides, a special form was prepared to collect information from technical diploma and degree holders. Part - I of the household schedule was related to the agricultural holding of the household and the household industry, if any and Part - II was meant to give information about the members of the household.

A new occupational classification was made in this census. The whole population was divided into two categories namely, working class and non-working class.

Census 1971 :

The special features of 1971 census were,

(i) the data were collected on current fertility.

(ii) a separate card known as, 'degree holders and technical personnel card' was filled in.

(iii) for the first time, detailed information was collected on migration.

(iv) for the first time, census data were processed by using computers.

The 1971 census also covered establishments where retail or wholesale business were carried on, or commercial services were rendered, or an office (public or private) or a place of entertainment or a place where educational, religious or social services were rendered. It was necessary that in each of such places one or more persons should actually be working.

Census 1981 :

Two schedules were canvassed this time. Like 1971 census, one schedule related to household and the other, to the individual slip. The information about the household with regard to the number and distribution of family members, religion of the head of the family, whether scheduled caste or tribe, mother tongue, ownership of the house, availability of water and electricity, number of rooms in the house, land holdings etc. were recorded. The individual slip contained information on demographic data like relationship to the head of the household, sex, age, birth place, marital status, sociological information like nationality, religion, education, mother tongue and economic data like occupation, business establishment, employment etc. Separate cards were filled in for the persons holding a recognised degree or diploma in science, engineering, technology or medicine etc. This card contained information about the biodata and employment of the degree holder.

The Establishment schedule was replaced by the Enterprise List in the 1981 census. An enterprise is defined as an undertaking engaged in the production and / or distribution of goods and / or services not for the purpose of own consumption.

Census 1991 :

It was the fifth census of independent India. As before, the whole of the villages and wards of the country were divided into enumeration blocks in such a way that in rural area an enumeration block consisted of around 750 people and in urban area 600 people. Along with houselisting operation which costituted the first phase of census 1991, the Central Statistical Organisation (CSO) of India concurrently canvassed an enterprise list as a part of the Third Economic Census. Each financial undertaking had to fill up an enterprise list furnishing information with regard to the nature of activity, classification, type of ownership, number of workers employed, types of energy used by the enterprise etc.

In the census 1991, the information were collected through the family schedule having 34 columns. Columns 1-7 contained the information like family serial number, name of the person, relation to the head of family, sex, age, mother language and marital status.

The individual slip canvassed related to the information about an individual. Questions 1-13 were of general nature i.e., with regard to name, relation with head of the family, sex, age, marital status, mother tongue and any other two languages known to him/her, religion, caste, literate or illiterate, educational standard, whether attending a school or college or not. Previously children in the age group 0-6 were classified as illiterate (it is immaterial whether they learnt reading or writing). In census 1991, this age group was fixed at 0-4. Three questions 14-16 related to the detailed enquiry about economic activity. Question 17 was specially included for retired military personnel to know whether they are getting pension or not. Questions 18-20 were in respect of place of birth, migration and reason (s) for migration. The 21st question was with regard to the length of stay at the place of enquiry at the time of enumeration. The 22nd question related to only those women who were not married, divorced or widow. The last question No - 23 related to married women asking whether they gave live birth during last one year or not.

Census 2001 :

According to 2001 census, Odisha had a population of 3,68,04,660 of which 1,86,60,570 are males and 1,81,44,090 are females. This is 3.64 percent of India's total population of 1, 02,86,10,328 recorded in the 2001 census. Odisha continues to maintain the 11th position in the ranking among the States and Union Territories of India so far as the population size is concerned.

The density of population (per sq.km) of Odisha was 113 persons in 1961 which increased to 141 during 1971, to 169 during 1981. According to 1991 census, density of population was 203 which has further increased to 236 in 2001 census. The density of population per sq.km for the country as a whole is 324 according to the 2001 census.

In Odisha, 22.54 percent of the total population of 28,18,455 of 15 towns / cities stay in slums while 27.57 percent of the child population in the age group of 0- 6 years live in slums.

The sex-ratio of Odisha was maintained with surplus females over males upto 1961 i.e., 1000:1001 in 1961. However, a trend of persistent decline in the number of females is being recorded since the year 1921. Of the total literates, about one fifth (i.e.,19.21%) live in slums. While the sex-ratio among the total population in respect of the 15 towns is 1000:875, the sex-ratio of the slum population is a little higher i.e. 1000 : 910 The table below gives some details about the 2001 census.

Table

Basic Indicators

Sl.No.	Indicators	Unit	Odisha	India
1.	Total population	Number	3,68,04,660	1,02,86,10,328
	a. Males	Number	1,86,60,570	53,21,56,772
	b. Females	Number	1,81,44,090	49,64,53,556
	c. Rural	Number	3,12,87,422	74,24,90,639
	d. Urban	Number	55,17,238	28,61,19,689
2.	Caste			
	(a) SC			
	(i) Males	Number	30,73,278	8,60,88,760
	(ii) Females	Number	30,08,785	8,05,46,940
	(b) ST			
	(i) Males	Number	40,66,783	4,26,40,829
	(ii) Females	Number	40,78,298	4,16,85,411
3.	Literacy			
	(i) Males	Number	1,19,92,333	33,65,33,716
	(ii) Females	Number	78,44,722	22,41,54,081
4.	Rural population	Percent	85.03	72.22
5.	Urban population	Percent	14.97	27.78
6.	Rural decadal growth	Percent	13.80	17.94
7.	Urban decadal growth	Percent	29.78	31.17
8.	Population density (per sq km)	Number	236	324
9.	Population growth rate	Percent	15.94	21.34
10.	Sex ratio	Number	972	933
11.	Literacy	Percent	63.61	65.38

Source : Census Report of India, 2001.

***

EXERCISES

1. Name the important statistical organisations in india and indicate their functions.
2. What is the necessity of statistical organisations ? Name three important statistical organisations in India under the central Government. Write their functions and indicate two publications of each organisation.
3. What purpose is served by the creation of National Sample Survey Organisation ? Write its functions.
4. Explain in brief, the functions performed by :
   (a) Directorate of Economics and Statistics
   (b) Department of Commercial Intelligence and Statistics
   (c) Department of Research and Statistics (Reserve Bank of India).
5. What are the function of Central Statistical Organisation ? Give your suggestions to make it more effective and useful.
6. Describe the formation and functions of the Indian Statistical Institute.
7. Describe the procedure of obtaining yield statistics.
8. Describe two common defects of Indian Agricultural Statistics and suggest methods for their improvement.
9. What do you understand by crop cutting experiments ? What is the purpose of such experiments ?
10. Describe the Annawari method and Random sampling method of obtaining yield statistics. Which of the two is better ?
11. What are the different methods of population census in India ? Describe each method in detail.
12. Distinguish between de facto and de jure methods of population census. Which method of census is used now a days in India ?
13. Describe the special features of the last census of population in India.
14. Give an account of the method of crop estimation followed in India.
15. What is the purpose of population census in India ? Describe in brief how population census is held in India.
16. Name a few statistical organisations in Odisha. Describe their functions and usefulnes.

17. Expand the following abbrevations.

(a) DES (b) ICAR (c) RGI (d) DCIS

(e) NSS (f) NSSO (g) CSO (h) DIS

18. (a) When was the first census held in India ?

(b) What is the time gap between two consecutive population censuses ?

(c) Who conducts the population census in India.

(d) When was the first complete population census held in India ?

(e) When will the next population census be held in India ?

(f) Name two publications of C S O.

(g) According to which committee recommendation, statistical bureaus were established in most of the states of India ?

(h) Name two publications in Agricultural Statistics.

(i) What is a cut in agricultural experiment as defined in random sampling method of crop estimation ?

(j) Who conducts the crop estimation in Odisha

(k) According to the 2001 population census held in India what was the total population ?

(l) What is the sex ratio in India as per 2001 population census ?

(m) What is the population growth rate as per 2001 census ?

(n) As per 2001 census, what are the total number of males and females in Odisha ?

(o) What is the density of population in India ?

* * *

MODEL QUESTIONS

1. **Which of the following is true ?**

(a) Official statistics are needed for

(i) improving the educational system in the state.

(ii) framing suitable national policies.

(iii) improving the infrastructure of industries.

(iv) developing the agricultural system prevailing in the state.

(b) To improve the agricultural statistics in India, Agriculture and Commerce department was set up in the year 1875 in which of the following states ?

(i) Bihar (ii) Odisha (Orissa) (iii) Uttar Pradesh (iv) West Bengal

(c) The first complete population census covering the whole of India was conducted in which of the following years ?

(i) 1881 (ii) 1885 (iii) 1872 (iv) 1891

(d) In order to enquire into the adequacy of the statistical data available and to recommend for their improvement, a committee under the chairmanship of which of the following persons was constituted ?

(i) P.C. Mahalonobis (ii) C.R. Rao (iii) Sir. R.A.Fisher

(iv) Sir. M.Visweswaraya

(e) Which of the following acts as the advisory and coordinating body for statistical system in India ?

(i) C.S.O (ii) N.S.S.O (iii) I.S.I (iv) I.C.A.R

(f) Who of the following is responsible for collection, compilation and publication of agricultural statistics in India ?

(i) Directorate of Economics and Statistics

(ii) Institute of Agricultural Research Statistics

(iii) National Sample Survey Organisation

(iv) Central Statistical Organisation

(g) Which of the following does not come under agricultural statistics ?

(i) Land utilisation (ii) Agricultural wages and prices

(iii) Statistics regarding livestock, poultry, forestry and fisheries

(iv) Statistics of labours engaged in industries

(h) Which of the following is not included under food crop estimation ?

(i) Pulses (ii) Spices (iii) Oats (iv) fruits

(i) Which of the following is not true ?

(i) Population statistics are a part of the science of demography.

(ii) There are two methods of population census viz. De facto method and De jure method.

(iii) In De facto method, people are counted on the basis of their habitual residence.

(iv) The first census in free India was conducted in the year 1951.

(j) For estimation of yield of crops, which of the following methods is used ?

(i) Simple Ramdom Sampling (ii) Stratified Multistage Random Sampling

(iii) Systematic sampling (iv) Stratified Random Sampling

2. **(a) Fill in the blanks.**

(i) One of the functions of CSO is to coordinate statistical activities between _____ and ______.

(ii) The National Sample Surveys were conducted for the first time in ______.

(iii) Technical guidance for corp estimation surveys and crop cutting experiments are provided by ______ in Odisha.

(iv) Indian statistical Institute brings out a statistical journal of international repute named ______.

(v) To collect land use statistics ______ method is followed in Odisha.

(b) Indicate True (T) or False (F) in the following :

(i) Now a days crop cutting experiments are conducted by the DES and statistical section of the Board of Revenue.

(ii) One of the disadvantages of stratified multistage random sampling method is that the standard error of the estimate can not be measured.

(iii) Population census in India is organised by a non-goverment organisation with the cooperation of Primary and High school teachers.

(iv) In De jure method, people are counted wheresever they are present on the census night.

(v) The next population census in India will be held in 2021.

3. **Give short answers to the following questions :**

(i) Indicate two functions of CSO

(ii) Mention two publications on area and yield of crops.

(iii) Indicate two functions of Directorate of Economics and Statistics in Odisha.

(iv) What purpose is served by land utilisation statistics ?

(v) Indicate two main functions of IARS.

(vi) What do you mean by agricultural statistics ?

(vii) Indicate the advantages of De-facto method of population census.

(viii)Write the disadvantages of De-jure method of population census.

(ix) What is population census ? Which method was followed in the population census, held last, in India ?

(x) Write the purpose of population census.

ANSWERS

1 (a)(ii) (b) (iii) (c) (i) (d) (iv) (e) (i)
(f) (i) (g) (iv) (h) (iii) (i) (iii) (j) (ii)

2. (a) (i) Centre, State (ii) 1950 (iii) ISI (iv) Sankhya
(v) Stratified multistage random sampling

(b) (i) T (ii) F (iii) F (iv) F (v) T

***