STATISTICAL ANALYSIS

Lessons 1 to 14

Dr. Kuldeep Kumar Attri
Dr. Suresh Sharma

International Centre for Distance Education & Open Learning
Himachal Pradesh University,
Summer Hill, Shimla - 171005

Contents

MC 1.4
STATISTICALANALYSIS FOR DECISION MAKING

Max Marks 80
Internal Assessment 20

Note: There will be Ten (10) questions in the paper spread into Five Units as Two questions from each unit. The candidate will require to attempt one question from each unit. Each question will carry Sixteen (16) marks.

COURSE OBJECTIVE:

The objective of this course is to provide an in-depth understanding of basic theoretical and applied principles of statistics needed to enter the job force. Students will be able to communicate key statistical concepts to non-statisticians. Students will gain proficiency in using statistical software for data analysis.

COURSE CONTENTS:

UNIT-I

Measurements of central tendency, dispersion, skewness and kurtosis.

UNIT-II

Regression analysis and correlation analysis (Two variables only).

Index Numbers: Meaning, construction of index numbers, problem in the construction of indexnumbers. Price, Quantity and Value Indices.

UNIT-III

Probability Theory: Probability, Classical Probability, Relative frequency Probability and Subjective Probability. Addition and multiple theorems of probability and Bay's Theorem. Probability distribution. Binomial distribution. The Poisson distribution and the Normal distribution.

UNIT-IV

Statistical Inferences; Testing of Hypotheses and Estimation, Sampling Distributions and Procedure of Testing Hypotheses Hypothesis Testing: Large and small sample tests (Z test, T test)

UNIT –V

F-test and Non-Parametric Test: Chi-square, run test. Sign test, Median test. Rank Correlation test, Kruskal-Wallis Test

COURSE OUTCOME:

Upon completion of the program, students will be able to:-

%4 u Demonstrate knowledge of probability and the standard statistical distributions.

%4 u Demonstrate knowledge of fixed-sample and large-sample statistical properties of point and interval estimators.

%4 u Demonstrate knowledge of the properties of parametric, semi-parametric and nonparametric testing procedures.

%4 u Demonstrate the ability to perform complex data management and analysis.

%4 u Demonstrate understanding of how to design experiments and surveys for efficiency.

References:

Johnson, R.D and Siskin, B.R Quantitative techniques foe business decision. Prentice Hall of India, 1984. Hien, L.W- Quantitative Approach to managerial decision. Practice Hall of India, 1983.

Levin, Richard I. and Rubin David S - Statistics for management, Prentice Hall of India, 1983. Chou- Ya- Lun; Statistical Analysis. Holt, Rinehart and Winston, 1980.

Fruend, J.E and William. F.J Elementary Business Statistics - The Modern Approach, 1982

Hooda, R.P, Statistical Methods.

*****

Lesson-1
Measures of Central Tendency

Structure:

1.1. Learning Objectives: After studying the lesson, you should be able to understand:

Meaning and objects of Measures of Central Tendency.

Different properties of a good average.

Different methods of measuring Central Tendency.

Introduction:

One of the important objects of statistical analysis is to find out various numerical measures which explains the inherent characteristics of a frequency distributions. The first of such measures is averages. The averages are the measures which condense a huge unwieldy set of numerical data into single numerical values which are representative of the entire distribution. The inherent inability of the human mind to grasp in its entirely a large body of numerical data compels us to seek relatively few constants that will describe the data. Averages provide us the gist and give a bird's eye view of the huge mass of unwieldy numerical data. Averages are the typical values around which other items of the distribution congregate. They are the values which lie between the two extreme observation of the distribution and give us an idea about the concentration of the values in the central part of the distribution. They are sometimes called as the measures of Central Tendency.

Averages are also called Measures of location since they enable us to locate the position or place of the distribution in question. Averages are statistical constants which enables us to comprehend in a single effort the significance of the whole. In other words, an average is a single value selected from a group of values to represent them in some way—a value which is supposed to stand for whole group of which it is a part. According to Croxton and Cowden, an average value is a single value within the range of the data, it is sometimes called a measure of central value.

To conclude, an average known as the measure of central tendency is the most typical representative item of the group to which it belongs and which is capable of revealing all the important characteristics of that group or distribution.

1.3. Objects of Measures of Central Tendency:

The most important object of calculating an average or measuring central tendency is, to determine a single figure which may be used to represent a whole series involving magnitudes of the same variable.

The second object of calculating an average represents the entire data it facilitates comparison within one group or between group of data. Thus the performance of the members of a group can be compared by relating it to the average performance of the group.

The third object is that an average helps in computing various other statistical measures such as Dispersion, Skewness, Kurtosis etc.

1.4. Essentials of a Good Average:

Since an average represents the statistical data and is used for purpose of comparison it must posses the following properties:

It must be rigidly defined and not left to the more estimation of the observer. If the definition is rigid the computed value of the average as obtained by different persons shall be alike.

The average must be based upon all values given in the distribution. If the items are not so based it might not be representative of the entire data.

It should be easily understandable. This would be so if it possesses some simple and obvious properties. It should not be too abstract for the common people.

It should be capable of being calculated with reasonable care and rapidity.

It should be least affected by fluctuations of sample. In other words the method should be such that the average computed from various samples of given data vary by least possible amounts.

It should be such that it can lend itself readily to algebraically treatment. In other words it may be possible to determine the average of a series by the use of the averages of its component parts. Different methods of measuring "Central Tendency" provide us with different kinds of average. The following are the main types of averages that are commonly used:

1 . Arithmetic Mean
Median
Mode
Geometric Mean and
Harmonic Mean

(A) Self Assessment

Fill in the blanks:

1. **The huge mass of unwieldy data is summarized in the form of .... and ....**
2. **A .... or an average is very essential and an important summary measure in any Statistical analysis.**
3. **An ....   is a single value which can be taken as representative of the whole distribution.**

A measure of central tendency is a typical value around which other figures ....

1.5. Arithmetic Mean

The arithmetic mean of a series is the quotieni obtained by dividing the sum of the values by the number of items.

In algebraic language, if X_1, X_2, X_3 .... + X_n are the N values of a variate X, then the Arithmetic Mean $\bar{X}$ is defined as follows :

$$\bar{x} = \frac{1}{N}\ (X_1\ X_2\ X_3.....X_n \quad = \frac{1}{N}(\ \ X) = \frac{X}{N}$$

Example 1. The following are the monthly salaries in rupees of ten employees in an office. Calculate the mean salary of the employees.

250,   275   265,   280   400, 490,   670,   890,   1100,   1250.

Solution:

$$\overline{X} = \frac{X}{N} = \left[\frac{250+275+265+280+400+490 +670+890+1100+1250}{10}\right] = \frac{5870}{10} = \text{Rs.587.}$$

Short-cut method: The example 1 was solved by direct method. The direct method is suitable where the number of items is moderate and the figures are small sized and integers. But if the number of items is large and/ or the values of the variate are big, then the process of adding together all the values may be a lengthy one. To overcome this difficulty in computations, a short-cut method may be used. This short cut method of computation is based on an important characteristic of the arithmetic mean, that is, the algebraic sum of the deviations of a series of individual observations, from their mean is always equal to zero. Thus deviations of the various values of the variate from an assumed mean are computed and the sum is divided by the number of items. The quotient so obtained is added to the assumed mean. The resultant figures is the arithmetic mean. Symbolically,

$$\overline{X}=A \qquad \frac{X}{N}$$

Where A = assumed mean and deviation or d = (X - A)

Example 2. Solve Example 1 by short cut method.

Solution :

Computation of Arithmetic Mean (Short-cut Method)

Serial Number	Salary in Rupees X	Deviations from Assumed Mean, A = 400, d = X-A
1	250	-150
2	275	-125
3	265	-135
4	280	-120
5	400	0
6	490	+ 90
7	670	+ 270
8	890	+ 490
9	1100	+ 700
10	1250	+ 850
N=10		d = + 2400 - 530 = + 1870

$$\overline{X} = A \quad \frac{X}{N} \quad 400 \quad \frac{1870}{10} \quad 400 \quad 187 \quad 587 \text{ (Rupees)}$$

Calculation of Arithmetic Mean in Discrete Series. In discrete series also arithmetic mean may be computed by both direct and short cut methods. The formula according to direct method is:

$$\overline{X} = \frac{1}{N} \quad (f\,X_{1}\ fX_{12}\ {}_{2}\ldots\ldots\ f\,X_{n\ n}) \quad \frac{(fX)}{N} \ .$$

where the variable values X_1, X_2 .... X_n, have frequencies f_1, f_1, ... f_n, and N = f.

Example 3. The following table gives the distribution of 98 accidents during seven days of the week of a given month. During the particular month there were 5 Fridays and Saturdays and only four each of other days. Calculate the number of accidents per day.

Days	No. of accidents
Sunday	20
Monday	22
Tuesday	10
Wednesday	9
Thursday	9
Friday	8
Saturday	20
Total	98

Solution:

Calculation of Number of Accidnts Per day

Day	No. of accidents X	No. of days in month f	Total accidents fx
Sunday	20	4	80
Monday	22	4	88
Tuesday	10	4	40
Wednesday	9	4	36
Thursday	9	4	36
Friday	8	5	40
Saturday	20	5	100
	98	N=30	420

$$\bar{X} \quad \frac{(fX)}{N} \quad \frac{420}{30} \quad \text{14.5 accidents per day}$$

The formula for computation of arithmetic mean according to the SHORT CUT METHOD is

$$\bar{X} \quad A \quad \frac{(fX)}{N}$$

where A = Assumed mean
d = X-A
N = f.

Example 4. Compute the arithmetic mean of the data given in Example 3 by short-cut method.

Solution: Calculation of Accidents per day by Short Cut Method

Day	X	d = X-A where A = 10	f	fd
Sunday	20	+ 10	4	+ 40
Monday	22	+ 12	4	+ 48
Tuesday	10	+ 0	4	+ 0
Wednesday	9	- 1	4	- 4
Thursday	9	- 1	4	- 4
Friday	8	-2	5	- 10
Saturday	20	+ 10	5	+ 50
			30	+ 120

$\bar{X}$ A $N^{\underline{fd}}$ 10 $\frac{120}{30}$ 10 4 14 accidents per day

Calculations of Arithmetic Mean for Continuous Series: Here also, the arithmetic mean can be computed both by direct and short-cut method. In addition, sometimes, a coding method or step deviation method is also applied for simplification of calculations. In any case, it is necessary to find out the mid-values classes in frequency distribution before arithmetic mean of the frequency distribution can be computed. Once the mid-points of the various classes are found out, then the process of the calculation of arithmetic mean is same as in the case of discrete series. In the case of direct method, the formula is:

$$\_\ \ \underline{fm}$$

$$N$$

where m = mid-points of various classes
N = the total frequency
In the short method, the following formula is applied

$$\bar{X}\ A\ \ \frac{fd}{N}\ \text{where d= (m - A) N = } \ \text{f.}$$

The short-cut method can further be simplified in practice by what is known as coding method. The deviations from the assumed mean are divided by a common factor to reduce their size. The sum of the products of the deviations and frequencies is multiplied by this common factor and it is divided by the total frequency and added to the assumed mean. Symbolically

$$\bar{X}\ A\ \ \frac{fd'}{N}\ \text{c.i.}$$

where

$$d\ \ \frac{m\ \ A}{c.i}\text{; c.i. = common interval or factor.}$$

Examples 5: Following is the freqency distribution of marks obtained by 50 students in a certain test in statistics:

Marks	Number of Students
0-10	4
10-20	6
20-30	20
30-40	10
40-50	1
50-60	7

Calculate arithmetic mean by
direct, (ii) short-cut and (iii) coding
methods. Direct Method :

Solution: Calculation of Arithmetic Mean

X	f	m	fm	d = m - A where A=25	$d'\ \frac{m\ \ A}{c.i}$ where c.i. = 10	fd	fd’
0 10	4	5	20	20	2	80	8
10 20	6	15	90	10	1	60	6
20 30	20	25	500	0	0	0	0
30 40	10	35	350	+10	+1	+100	+10
40 50	7	45	315	+20	+2	+140	+14
50 60	3	55	165	+30	+3	+90	+9
	N=50	fm = 1440			fd = 190fd’= + 19		

$$\bar{X}\quad \frac{fm}{N}\quad \frac{1440}{50}\ 28.8 \text{ Marks.}$$

Short-cut Method:

$$\bar{X}\ A\quad \frac{fd}{N}\ 25\quad \frac{190}{50}\ 25\ \ 3.8\ \ 28.8 \text{ Marks.}$$

Coding Method:

$$\bar{X}\ A\quad \frac{fd'}{N}\ c.i.\ 25\quad \frac{19}{50}\ 10\ \ 25\ \ 0.38\ \ 10\quad 25\ \ 3.8\ \ 28.8 \text{ Marks.}$$

None that answer is same in the three cases.

Calculations of Combined Arithmetic Mean : If a series of N observations of a variable X consists of two component series, the Arithmetic mean of the whole series can be readily expressed in terms of the means of the two component series. If, for example, we denote the values in the first component series by X_1 and the second series by X_2, then

$$(X) = \quad (X_1) + \quad (X_2)$$

Further, if there are N_1 observations in X_1 series and N_2 in X_2 series, then the mean of the combined series will be equal to the means of X_1 and X_2 series by the following formula:

$$\bar{X}\quad \frac{N_1\bar{X}_1\ \ N_2\bar{X}_2}{N_1\ \ N_2}$$

where X_1, represents combined arithmetic mean N_1 and N_2 represents the frequencies of two series.

Example 6. The mean age of the group of men is 32 and that of the group of women is 27. The total number of men and women are 60 and 40 respectively. Find the combined mean of both the groups.

Solution. Given

$$\bar{X}_m\quad 32 \text{ years and } N_m\quad 60$$

$$\bar{X}_w\quad 27 \text{ years and } N_w\quad 40$$

By putting the values in the formula of the combined mean, we get

$$\bar{X}_{mw}\quad \frac{32(60)\ \ 27(40)}{60\ \ 40}\quad \frac{1920\ \ 1080}{100}\quad \frac{3000}{100}\ 30 \text{ years.}$$

Correcting Incorrect Value of Mean:

It sometimes happens that due to an oversight or mistake in copying, some wrong items are taken while calculat-ing mean. Then the problem arises how to correct mean so calculated. It is very simple. Find out incorrect X by the formula:

$X=N\bar{X}$

Then deduct wrong items from incorrect X and add correct items and divide the corrected X by number of observations. The correct mean will be:

$$\text{Correct mean}\,\bar{X} \quad \frac{\text{Correct } \; X}{N}$$

Example 7. The average marks secured by 100 students was 88. Later on, it was discovered that a score of 36 was misread as 63. Calculate the correct average marks secured by the students.

Solution. Given

N =100

$\bar{X} \quad 8\underline{8}$

$X^{\underline{X}\,N}$

Therefore $X \quad N\bar{X}$

$X = 100 \quad 80 = 8800$

Correct $\quad X \quad$ Incorrect X - Wrong item

Correct item

$8800 \;\; 63 \;\; 36 \;\; 8827$

$$\text{Correct}\,\bar{X} \quad \frac{\text{Correct } \; X}{N} \quad \frac{8827}{100}\,88.27$$

Therefore the correct average is 88.27.

Calculation of Arithmetic mean in Case of Open-End Classes:

Open-end classes are those in which lower limit of the first class and the upper limit of the last class are not know. In these kinds of series we can not calculate mean unless we make an assumption about the unknown limits. The assumption depends upon the class-interval following the first class and preceding the last class. For example:

Marks	No. of students
Below 15	4
15-30	6
30-45	12
45-60	8
Above 60	7

In this example because all the class-intervals are same, the assumption would be that the lower limit of the first class is zero and upper limit of last class is 75. Hence first class would be 0 — 15 and the last class 60 — 75.

What happens in his case ?

Marks	No. of students
Below 10	4
10-30	7
30-60	10
60-100	8
Above 100	7

In this problem because the class interval is 20 in the second class, 30 in the third, 40 in the fourth class and so on. The class interval is increasing by 10. Therefore the appropriate assumption in this case would be that the lower limit of the first class is zero and the upper limit of the last class is 150.

If the class intervals are of varying width, an effort should be made to avoid calculating mean and mode. It is advisable to calculate median.

The Weighted Arithmetic Mean:

In the computation of arithmetic mean we had given equal importance to each item in the series. This equal importance may be mislead if the individual values constituting the series have different importance as in the following illustration:

The Raja Toy Shop sells

Toy Cars at	**Rs. 3 each**
Toy Locomotives at	**Rs. 5 each**
Toy aeroplane at	**Rs. 7 each**

Toy Double Decker at Rs. 9 each

What shall be the average price of the toys sold. If the shop sells 4 toys one of each kind.

Mean Price i.e.

$$\underline{X} \qquad \frac{X}{N}$$

$$= Rs. \frac{24}{4} = Rs.6.$$

In this case the importance of each toy is equal in as much as one toy of each variety has been sold. In the above computation of arithmetic mean this fact has been taken care of by including once only the price of each toy. But if the shop sells 100 toys, 50 cars, 25 locomotives, 15 aeroplanes and 10 double deckers, the importance of the four toys to the dealer is not equal as a source of earning revenue. In fact their respective importance is equal to the number of units of each toy sold, i.e.

the importance of Toy car is	**50**
the importance of Locomotive is	**25**
the importance of Aeroplane is 15	
the importance of Double Decker is	**10**

It may be noted that 50, 25,15, 10 are the quantities of the various classes of toys sold. It is for these quantities that the term 'weights' is used in statistical language. Weight is represented by symbol W and W represents the sum of weights.

While determining the average price of toy sold these weights are of very great importance and are taken into account in the manner illustrated below:

$$\bar{X} \quad \frac{[(W_1 X_1)\ \ (W_2 X_2)\ \ (W_3 X_3)\ \ (W_4 X_4)]}{W_1 + X_2 W_3 + X_4}$$

$$\frac{WX}{W}$$

where W_1, W_2, W_3, W_4 are the respective weights and X_1, X_2, X_3, X_4 represent the price of 4 varieties of toy.

Hence by substituting the values of W_1, W_2, W_3, W_4 and X_1, X_2, X_3, X_4 we get

$$\bar{X} \quad \frac{(50\ \ 3)\ \ (25\ \ 5)\ \ (15\ \ 7)\ \ (10\ \ 9)}{50\ \ 25\ \ 15\ \ 10} \quad \frac{150\ \ 125\ \ 105\ \ 90}{100} \quad \frac{470}{100}\ \text{Rs.4.70}$$

The table given below summaries the steps taken in the computation of the Weighted Arithmetic Mean. Weignted Arithmetic mean of Toys by the Raja Shop

Toy	Price Per toy Rs. X	Number sold W	Price x weight WX
Car	3	50	150
Locomotives	5	25	125
Aeroplane	7	15	105
Double Decker	9	10	90
		W=100	WX=470

$$\bar{X} \quad \frac{WX}{W} \quad \frac{470}{100} \quad \text{Rs.4.70}$$

Example 8. The table below shows the number of skilled and unskilled workers in two localities, together with their average hourly wages.

Determine the average hourly wage in each locality. Also give reasons why the results show that the average hourly wage in Shyam Nagar exceed the average hourly wage in Ram Nagar, even though in Shyam Nagar the average hourly wages of both categories of workers is lower.

Ram Nagar		
Workers Category	No.	Wages (per hour)
Skilled	150	1.80
Unskilled	850	1.30
Shyam Nagar		
Workers Category	No.	Wages (per hour)
Skilled	350	1.75
Unskilled	650	1.25

It is required to compute weighted arithmetic mean.

Solution :

	Ram Nagar				Shyam Nagar		
	X	W	WX		X	W	WX
Skilled	1.80	150	270	Skilled	1.75	350	621.50
Unskilled	1.30	850	1105	Unskilled	1.25	650	812.50
		1000	1375			1000	1425
	$\overline{X}_w \quad \frac{1375}{1000}$ = Rs. 1.375					$\overline{X}_w \quad \frac{1425}{1000}$ = Rs. 1.425	

It may be noted that weights are more evenly assigned to the different categories of workers in Shyam Nagar than in Ram Nagar.

1.6. Median

The median is that value of the variable which divides the group in two equal parts, one part comprising the value greater and the other all values less than median. Median of a distribution may be defined as that value of the variable which exceeds and is exceeded by the same number of observation. It is the value such that the number of observations above it is equal to the number of observations below it. Thus, we come to know that the arithmetic mean is based on all items of the distribution, the median is only positional average, that is, its value depends upon the position occupied by a value in the frequency distribution.

When the items of a series are arranged in ascending or descending order of magnitude the value of the middle item in the series is known as median in the case of individual observations. Symbolically:

$$\text{Median = Size of} \frac{N \quad 1}{2} \text{ th item}$$

If the number of items is even, then there is no actual value exactly in the middle of the series. In such a situation the median is arbitrarily taken to be halfway between the two middle item. Symbolically:

$$\text{Median} = \frac{\text{Size of } \frac{N}{2} \text{ th item} + \frac{N+1}{2} \text{ th item}}{2}$$

Example 9. Find the median of the following two series:

(i) 8, 4, 8, 3, 4, 8, 6, 5, 10

Solution :

Computation of Median

(i)		(ii)	
S.No.	X	S. No.	X
1	3	1	5
2	4	2	5
3	4	3	7
4	5	4	9
5	6	5	11
6	8	6	12
7	8	7	15
8	8	8	28
N = 9		N = 8	

15,12, 5, 7, 9, 5, 11, 28

In this case of (i) Series,

Median = Size of $\frac{N+1}{2}$ th item = Size of $\frac{9+1}{2}$ th item = size of 5th item = 6

In this case of (ii) series,

Median = Size of the $\frac{N+1}{2}$ th item = Size of the $\frac{8+1}{2}$ th item

$= \frac{\text{Size of 4th item + Size of 5th item}}{2}$ = Size of 4.5th item $\frac{9+11}{2}$ 10.

Location of Median to Discrete Series: In the discrete series, the total frequency is divided into two equal parts. For this purpose cumulative frequency is found out.

Then size of the $\frac{N+1}{2}$ th item is located.

Example 10. The table-below shows the number of rooms in the houses of a particular locality. Find median of the data:

No. of rooms:	3	4	5	6	7	8
No. of houses:	38	654	311	42	12	2

Solution :

Computatin of Median

No. of Rooms	No. of Houses	Cumulative Frequency
X	f	c.f
3	38	38
4	654	692
5	311	1003
6	42	1045
7	12	1057
8	2	1059

Median = Size of the $\frac{N+1}{2}$ th item

= the size of $\frac{1059+1}{2}$ the item = the size of 530th item.

Median lies in the cumulative frequency of 692 and the value corresponding to this is 4. Therefore Median = 4 rooms.

Calculation of Median in a Continuous Series: The median is computed by any of the following formulae of interpolation:

$$\text{Med}\ l_1 \frac{\frac{N}{2}\,cf}{f_0}\ (l_2 - l_1)$$

where cf_o refers to the comulative frequency of the class preceeding to the median class.

$$\text{or } l_2 \quad \frac{\frac{N}{2} cf_0}{f} (l_2 - l_1)$$

Example 11. The following table gives you the distribution of marks secured by some students in an examination.

Marks	No. of Students
0 20	42
21 30	38
31 40	120
41 50	84
51 60	48
61 70	36
71 80	31

Find the median marks.

Solution :

Calculation of Median Marks

Marks (X)	No.of Students (f)	C.F.
0 20	42	42
21 30	38	80
31 40	120	200
41 50	84	284
51 60	48	332
61 70	36	368
71 80	31	399

Med. = Sizeof $\frac{N}{2}$ th item = Size of $\frac{399}{2}$ th item or 199.5 item

which lies in (31 - 40) group.

Hence median class is 30.5 — 40.5 Applying the formula of interpolation,

$$\text{Med } l_1 \quad \frac{\frac{N}{2} cf_0}{f} (l_2 - l_1) \quad 30.5 \ \frac{199.5 \ \ 80}{120}(10) \quad 30.5 \quad \frac{119.5}{120}(10)$$

$$30.5 \quad \frac{119.5}{120} \quad = 30.5 + 9.96 = 40.46 \text{ marks}$$

Note: The students should note that the actual limits of the group (31 — 40) are (30.5 — 40.5).

Related Positional Measures: The median divides the series into two equal parts. Similarly, there are certain other measures which divide the series into certain equal parts. There First quartile, third quartile, deciles percentiles etc. If the items are arranged in ascending or descending order of magnitude, Q_1 is that value which covers l/4th of the total number of items. Similarly, if the total number of items are divided into ten equal parts, then, there small be nine deciles.

Symbolically,
First quartile

(Q_1) = the value of $\dfrac{N+1}{4}$ th item

Third quartile

(Q_3) = the value of $\dfrac{3N+1}{4}$ th item

First decile

(D_1) = the value of $\dfrac{N+1}{10}$ th item

Sixth decile

(D_6) = the value of $\dfrac{6N+1}{10}$ th item

First percentile

(P_1) = the value of $\dfrac{N+1}{100}$ th item

Once values of the items are found out, then formulae of interpolation are applied for ascertaining the value of Q_1, Q_2, D_1 etc.

Example 12. Calculate Q_1, Q_3, D_2 and P_5 from the following data:

Marks	No. of Students
Below 10	8
10 20	10
20 40	22
40 60	25
60 80	10
above 80	5

Solution :

Calculation of Positional Values

Marks (X)	No.of Students (f)	C.F.
Below 10	8	8
10 20	10	18
20 40	22	40
40 60	22	65
60 80	10	75
above 80	5	80
	N=80	

Q_1 = size of $\frac{N}{4}$ the item $\frac{80}{4}$ = 20th item

Hence Q1 lies in the class 20 40

$$Q_1 \quad L_1 \quad \frac{\frac{N}{4}cf_0}{f} \; i$$

where L_1 = 20, $\frac{N}{4}$ = 20, c.f. = 18, f = 22 and i = 20

Q_1 =20+ $\frac{(20\text{-}18)}{22}$ 20 = 20 + 1.8 = 21.8.

In a similar way we can calculate

Q_3 = size of $\frac{3N}{4}$ th item = size of $\frac{3\ \ 80}{4}$ th item = 60th item

Hence Q_3 lies in the class 40 60

$$Q_1 \quad L \, \frac{\frac{3N}{4}\ cf0}{f} \quad i$$

Where L = 40

$\frac{3N}{4}$ 60, c.f., 40.

$$Q_3 \ 40 \quad \frac{(60\text{-}40)}{25} \ 20$$ = 40 + 16 = 56

D_2 = Size of $\frac{2N}{10}$ th item

D_2 = Size of $\frac{2\ \ 80}{10}$ 16th item

Hence D_2 lies in the class 10 20

$$D_2 \ L \quad \frac{\frac{2N}{10}\ cf0}{f} \quad i$$

where L = 10, $\frac{2N}{10}$ 16, c.f. 8. f = 10, i = 10

$$D_2 \quad 10 \quad \frac{16\text{-}8}{10} \qquad 1010818$$

P_5 = Size of $\frac{5N}{100}$ th item $\frac{5\ \ 80}{100}$ =4th item

Hence P_5 lies in the class 0 10

$$P_5 \quad L \frac{\frac{5N}{100}\ \text{c.f.}}{f} \quad i$$

where L = 0, $\frac{5N}{100}$ = 4, c.f. = 0, f = 8, i = 10

By substituting the values we get

$$P_5 \quad 0 \frac{4 - 0}{8} \; 10055.$$

Determination of Partition values in a continuous series: The formulae for the calculation of First and Third quartiles are obtained by replacing $\frac{N}{2}$ in the median formula $\frac{N}{4}$ by $\frac{3N}{4}$ and respectively.

Thus $$Q_1 \; L_1 \quad \frac{\frac{N}{4}\,cf0}{f} \; (L_2 \quad L_1)$$

$$Q_3 \quad L_1 \; \frac{\frac{3N}{4}\,cf0}{f} \; (L_2 \quad L_1)$$

Similar formula can be obtained for other partition values.

Example 13. Find the quartiles, P_{50} and 5th decile from the following data:

Marks	No. of Students
10 20	70
20 30	55
30 40	140
40 50	35
50 60	100
60 70	90
70 80	140
80 90	70

Solution.

The cumulative frequencies of the various classes are : 70, 125, 265, 300, 400, 490, 630 and 700. Applying the formula of integration,

$$Q_1 \; L_1 \quad \frac{\frac{N}{4}\,cf0}{f} \, (L_2 \; -L_1);$$

$$\frac{N}{4} \quad \frac{700}{4} \quad 175$$

$$3\,0 \qquad \frac{175 \;\; 125}{1\,4\,0} \; (40 \;\; 30)$$

$30 \quad 140^{50} (10) \quad 30 \quad 14\frac{50}{} \quad 30 \quad 3.57 \quad 33.57.$

$$Q_2 L_1 \quad \frac{\frac{3N}{4} cf_0}{f}(L_2 - L_1) \quad = 70+ \frac{525 \quad 490}{140}(10) \quad 70 \quad \frac{35}{140}(10);$$

where $\frac{3N}{4} \quad \frac{3 \quad 700}{4} \quad \frac{2100}{4} = 525$

$$70 \quad \frac{10}{4} \quad 70 \quad 2.5 \quad 72.5$$

$P_{50}\, 50 \quad \frac{350 \quad 300}{100}(10) \quad 50 \quad \frac{50}{100}(10)$; where $\frac{50\ N}{100} \quad \frac{50 \quad 700}{100} \quad \frac{3500}{100} \; 350 \quad 50555$

$D_5 \; L_1 \quad \frac{\frac{5N}{10} cf_0}{f}(L_2 - L_1)$; $=50 + \frac{350 \quad 300}{100} \quad (10)$; $\frac{5N}{10} \quad \frac{5(700)}{10} \quad 350$

$=50= \frac{50}{10} =50+5=55$

Note: That the value of P_{50} and D_5 will be same as median. Calculate Median and value thereof would also be 55.

$$\text{Med} = L_1 \quad \frac{\frac{N}{2} cf_0}{f}(L_2 \quad L_1); \quad \frac{N}{2} \quad 350$$

Calculations of Missing Frequencies :

Example 14. In the frequency distribution of 100 families given below; the number of families corresponding to expenditure groups 20 — 40 and 60 — 80 are missing from the table. However, the median is known to be 50. Find out the missing frequencies.

Expenditure:

0—2020—4040—6060—8080—100

No. of

families: 14 ? 27 ? 15

Solution :

Let the missing frequencies for the classes :

20 — 40 be x and

40 — 80 be y

Computation of Median

Expenditure (in Rupees)	No. of families (f)	c.f. Less than
0 20	14	14
20 40	x	14 + x
40 60	27	14 + 27 +x
60 80	y	41 + x + y
80 100	15	41 + 15 + x + y
	N=100	= 56 + x + y

From the table, we have

N = F = 56 + x + y = 100.

x + y = 100 56 = 44

Because median is given to be 50 which lies in the class 40 — 60 therefore 40 — 60 is the median class. By using the median formula we get :

$$\text{Median} = L + \frac{\frac{N}{2}\,cf_{\qquad 0}}{2}\ i\ ;\ 50 = 40 + \frac{50\ \ (14\ \ x)}{27}\quad (60\ \ 40)$$

$$-50=40+\ \frac{50\ \ (14\ \ x\,)}{27}\,20 \qquad \text{or } 50\ \ 40 \quad \frac{50\ \ 14\ \ x}{27} \quad 20$$

$$\text{or } 10\ \ \frac{36\ \ x}{27}\quad 20$$

$$\text{or } 10\quad (36\ \ x)\qquad \frac{20}{27}\ \ \text{or } 10\ \ 27 = 720 - 20\,x \text{ or } 270 = 720 -20\,x$$

$$20x = 720 -270 \qquad x = \frac{450}{20}\,22.5$$

By substituting the values of x in
x + y = 44
We get 22.5 + y = 44
= 44 22.5 = 21.5

Hence frequency for the class 20 40 is 22.5 and 60 80 is
21.5 1.7. Mode

The mode is that value of the variable which occurs or repeats itself maximum number of times. The mode is the most “fashionable” size in the sense that it is the most common and typical and is defined by Zizek as “the value occuring most frequency in series of items and around which the other items are distributed most densely.” In the words of Croxton and Cowden, the mode of a distribution is the value at the point where the items tend to be most heavily concentrated. According to A.M. Turtle, Mode is the value which has the greatest frequency density in its immediate neighbourhood. In the case of individual observations, the mode is that value which is repeated the maxi-mum number of times in the series. The value of mode can be denoted by the alphabet z also.

Example 1. Calculate mode from the following data :

S. No.	Marks Obtained
1	10
2	27
3	24
4	12
5	27
6	27
7	20
8	18
9	15
10	30

Solution:

Calculationof Mode Marks

X	f
10	1
12	1
15	1
18	1
20	1
24	1
27	3
30	1
	f = 10

Mode is 27 marks

Calculation of Mode is Discrete series. In this case also, it is quite often, possible to determine mode by inspection. Take in simple case :

X	f
1	4
2	5
3	13
4	6
5	12
6	8
7	6

By inspection, the Modal size is 3 as it has the greatest frequency. But this test of greatest frequency is not fool proof as it is not the frequency of a class, but also the frequencies of the neighbouring classes that decide the mode. In such cases, we have to resort to what is known as the method of grouping.

Example 2. Find out mode from the following data :

Size of Shoe	Frequency
1	4
2	5
3	13
4	6
5	12
6	8
7	6

Solution : By inspection, the mode is 3, but the size of mode may be 5. This is so because the neighbouring frequencies of size 5 are greater than the neighbouring frequencies of size 3. This effect of neighbouring frequencies is seen with the help of grouping frequencies in two's and three's.

Grouping Table

Size of	Frequency					
Shoe	1	2	3	4	5	6
	4					
		9				
2	5			22		
			18			
3	13				24	
		19				
4	6					31
			18			
5	12			26		
		20				
6	8				26	
			14			
	6					

Note : When there exist two groups of frequencies in equal magnitude then we should consider either both or omit both while analyzing the sizes of items.

Analysis Table

Columns	Size of items with greatest frequency
1	3
2	5,6
3	2,3,4,5
4	4,5,6
5	5,6,7
6	3,4,5

Thus 5 occurs greatest number of times, therefore mode is 5. It is to be noted that by inspection we found 3 to be the mode.

Determination of mode in continuous series :

In the continuous series, the determination of mode requires one step more than that for the discrete series. Once the modal class is determined by inspection or with the help of the process of group, then the following formula of interpolation is applied:

$$\text{Mode } l_1 \quad \frac{f_1 \quad f_0}{2f_1 f_0 \quad f_2}(l_2 \quad l_1)$$

$$\text{or } l_2 \quad \frac{f_1 \quad f_0}{2f_1 \quad f_0 \; f_2}(l_2 \, l_1)$$

l_1 = lower limit of the class where mode lies.

l_2 = upper limit of the class where mode lies.

f_0= frequency of the class preceding the modal class.

f_1 = frequency of the class where modal lies.

f_2 = frequency of the class succeeding the modal class.

Example 3. Calculate mode of the following frequency distribution:

Variable	Frequency
0 10	5
10 20	10
20 30	15
30 40	14
40 50	10
50 60	5
60 70	3

Grouping Table

Size of	Frequency					
Shoe	1	2	3	4	5	6
0 10	5					
		15				
10 20	10			30		
			25			
20 30	15				39	
		29				
30 40	14					39
			24			
40 50	10			29		
		15				
50 60	5				18	
			8			
60 70	3					

Analysis Table

Columns	Size of items with greatest frequency
1	20 30
2	20 30, 30 40
3	10 20, 20 30
4	0 10, 10 20, 20 30
5	10 20, 20 30, 30 40
6	20 30, 30 40, 40 50

Thus model group is (20—30) because it has occurred 6 times. Applying the formula of interpolation.

$$M_0 \quad l_1 \quad \frac{f_1 \quad f_0}{2f_1 \quad f_0 \, f_2}(l_2 \quad l_1) \quad 20 \quad \frac{15 \quad 10}{30\,10 \quad 14}(3020)$$

$$20 \quad \frac{5}{6}(10) \quad 20 \quad \frac{50}{6} \quad 28.3$$

Calculation of mode where it is ill defined. The above formula are not applied where there are many modal values in a series or a distribution. For instance there may be two or more than two having maximum frequency. In such a situation the series will be known as bimodal or multimodal. The mode is said to be ill-defined and in such cases the following formula of interpolation is applied.

Mode = 3 Median — 2 Mean.

Example 4. Calculate mode from the following data

Variable	Frequency
10 20	5
20 30	9
30 40	13
40 50	21
50 60	20
60 70	15
70 80	8
80 90	3

Solution. First of all, ascertain the modal group with the help of process of grouping.

Grouping Table

x	F					
	1	2	3	4	5	6
10 20	5					
		14				
20 30	9			27		
	22					
30 40	13				43	
		34				
40 50	21					54
			41			
50 60	20			56		
		35				
60 70	15				43	
			23			
70 80	8					26
80 90	3	11				

Analysis Table

Columns	Size of items with greatest frequency
1	40 50
2	50 60, 60 70
3	40 50, 50 60
4	40 50, 50 60, 60 70
5	20 30, 30 40, 40 50
	50 60, 60 70
6	30 40, 40 50, 50 60

There are two groups which occur equal number of items. They are (40—50) and (50—60). Therefore we will apply the following formula.

Mode = 3 Median — 2 Mean.

and for this purpose the values of mean and median are required to be computed.

Calculation of Mean and Median

Variate Value X	f	Mid Value m	$d' = \frac{m\ 45}{10}$	fd	cf
10 20	5	15	- 3	- 15	5
20 30	9	25	- 2	- 18	14
30 40	13	35	- 1	- 13	27
40 50	21	45	0	0	48 Median is the
50 60	20	55	+ 1	+ 20	68 value of $\frac{N}{2}$ th
60 70	15	65	+ 2	+ 30	83 item which lies
70 80	8	75	+ 3	+ 24	91 in (40 - 50) group
80 90	3	85	+ 4	+ 12	94
	N=94				fd' = + 40

$$\bar{X} \quad A \quad N\frac{fd'}{} \text{ c.i. } \quad 45 \quad \frac{40}{94}(10) \quad 45 \quad 4.2 \quad 45 \quad 49.2$$

$$\text{Median} = L_1 + \frac{\frac{N}{2}\ cf_0}{f} \text{ c.i.} = 40 + \frac{47\ \ 27}{21}(10) = 40 + \frac{200}{21} \quad 40 \quad 9.5 \quad 49.5.$$

Mode = 3 median - 2 mean = 3 (49.5) - 2 (49.2) = 148.5 - 98.4 = 50.1.

Determination of mode by curve fitting: The ideal method of calculating mode is curve fitting. For this purpose, the following steps are to be taken,

Draw a histogram of the data

Draw the lines diagonally inside the modal class rectangle, starting from each upper corner of the rectangle to the upper corner of the rectangle to the upper corner of the adjacent rectangle.

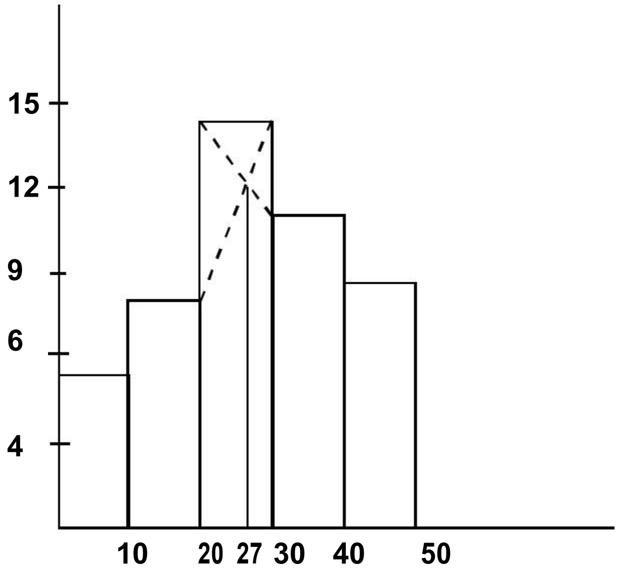
Draw a perpendicular line from the intersection of the two diagonal lines to the X-axis.

The abscissa of the point at which the perpendicular line meets is the value of the mode.

Example 5. Construct a histogram for the following distribution, determining the mode graphically:

Variable	Frequency
0 10	5
10 20	8
20 30	15
30 40	12
40 50	7

Verify the result with the help of interpolation.



$$\text{Mode}\ \ L_1\ \ \frac{f_1\ \ f_0}{2f_1\ \ f_0\ \ f_2}(L_2\ \ L_1)\ \ 20\ \ \ \frac{15\ \ 8}{30812}(30\ \ \ \ 20)\ \ 20\frac{7}{10}(10)20727$$

Example 6 :

Calculate mode from the following data.

Marks	No. of Students
Below 10	4
” 20	6
” 30	24
” 40	46
” 50	67
” 60	86
” 70	96
” 80	99

Solution : Since we are given the cumulative frequency distribution of marks, first we shall convert it into the frequency distribution as:

Marks	Frequencies
0 10	4
10 20	6 4 = 2
20 30	24 6 = 18
30 40	46 24 = 22
40 50	67 46 = 21
50 60	86 67 = 19
60 70	96 86 = 10
70 80	99 96 = 3
80 90	100 99 = 1

It is evident from the table that the distribution is irregular and maximum chances are that the distribution would be having more than one mode. You can test yourself by applying the grouping and analysing table.

The last and final way to calculate the value of mode in cases of bio-modal distribution is: Mode = 3 median - 2 mean.

Computation of Mean and Median

Variate	Mid Value (x)	Frequency (f)	c.f.	d = $\frac{X\ 45}{10}$	f.d.
0 10	5	4	4	-4	-16
10 20	15	2	6	-3	-6
20 30	25	18	24	-2	-36
30 40	35	22	46	-1	-22
40 50	45	21	67	0	0
50 60	55	19	86	1	19
60 70	65	10	96	2	20
70 80	75	3	99	3	9
80 90	85	1	100	4	4
		N = f = 100			f = - 28

$$\text{Mean}\ \ A\ \ \frac{fd}{N}\ i$$

$$\text{Mean}\ \ 45\ \ \frac{28}{100}\ 10$$

Mean 45 2.8 i

Mean 42.2

Here $\frac{N}{2}$ $\frac{100}{2}$ 50.

Because 50 is just short than 67 in c.f. column. Median Class is 40-59

$$\text{Median} = L + \frac{\frac{N}{2}\ cf}{f}\ i$$

$$\text{Median} = 40 + \frac{50\ \ 46}{21}\ 10$$

$$\text{Median} = 40 + \frac{4}{21}\ 10$$

Median 40 1.9 41.9

Hence Mode = 3 median - 2 mean Substitute mean and median values Mode = 3 41.9 - 2 42.2 = 125.7 - 84.3 = 41.3

(B) Self Assessment

State whether the following statements are true or false:

Arithmetic Mezh is defined as the sum of squares of observations divided by the number of observations.

In weighted arithmetic mean, the importance given to various observations is same.

The values of a variable that divide a distribution into four equal parts are called quartiles.

Deciles divide a distribution into 10 equal parts.

Percentiles divide a distribution into 100 equal parts :

Self Assessment

Multiple Choice Questions:

10. .... of distribution is that value of the variate which divides it into two equal parts.

Mean (b) Median (c) Mode (d) Standard deviation

Median is a... average because its value depends upon the position of an item and not on its magnitude.

Positional (b) Arithmetic (c) Mathematical (d) Commercial

Self Assessment

Fill in the blanks:

12. .... is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed.

For a moderately skewed distribution, the difference between mean and mode is approximately .... the difference between mean and median

1.8. Geometric Mean

We have so far discussed three measures of Central tendency—i.e. mean, median and mode. Now, we are taking up Geometric Mean. Both Geometric mean and Harmonic mean are used occasionally in business and economics.

In general, if we have n number (none of them being zero), then the G.M. is defined as

$$GM \quad \sqrt{x_1 x_2 \ldots\ldots x_n}$$

$$(x_1 x_2 \ldots\ldots\ldots \quad x_n)$$

In the case of a discrete series, if x_1, x_2 .... x_n occur f_1, f_2 .... f_n times respectively and N is the total frequency (i.e. N = f_1, f_2 .... f_n) then

$$GM = n\ x\ _1\sqrt{f_2 x_2 f_2 \ldots\ldots \quad x_n f_n}$$

For convenience, use of logarithms is made extensively to calculate the nth root. In terms or logarithms.

$$=GM = AL \frac{\log x_1 \quad \log x_2 \quad \ldots\ldots\ldots \ \log x_n}{N}$$

$$= AL \frac{\log x}{N}$$

where AL stand for anti log.

In discrete series,

$$GM=AL \frac{\underline{\log x}}{N}$$

and in the case of continuous series

$$GM=AL \frac{\underline{f \log m}}{N}$$

Example 7. Calculate G.M. of the following data :

Solution 2,4, 8

$$GM=3\sqrt{2}\sqrt{\quad 48} \qquad 3\sqrt{64 \quad 4}$$

In terms of logarithms, the question can be solved as follows:

log 2 = 0.3010

log 4 = 0.6921

log 8 = 0.9031

log x = 1.8062

Dividing it by 3 as N = 3, we get

$$\frac{\log x}{N} \quad \frac{1.8062}{3} = 0.60206$$

Finding Antilog, we get 4.

Thus G.M. = 4.

Example 8. Calculate geometric mean of the following data:

x	**5**	**67**	**8**	**9**	**10**	**11**
f	**2**	**4**	**7**	**10**	**96**	**2**

Solution.

Calculation of Geometric Mean

x	log x	f	f log x
5	0.6990	2	1.3980
6	07782	4	3.1128
7	0.8451	7	5.9157
8	09031	10	9.0310
9	09542	9	8.5878
10	1.0000	6	6.0000
11	1.0414	2	2.0828
	N=40		f log x = 36.1281

$$GM=AL \quad \frac{f \log x}{N} \quad AL \frac{36.1281}{40} = AL\ (0.9032) = 8.0020.$$

Example 9. Calculate G.M. from the following data:

Solution :

x	f
9.5 14.5	10
14.5 14.5	15
19.5 24.5	17
24.5 29.5	25
29.5 34.5	18
34.5 39.5	12
39.5 44.5	8

Solution.

x	m	log m	f	f log m
9.5 14.5	12	0.0792	10	10.7920
14.5 14.5	17	1.2304	15	18.4560
19.5 24.5	22	1.3423	17	22.8208
24.5 29.5	27	1.4314	25	35.7850
29.5 34.5	32	1.5051	18	27.0918
34.5 39.5	37	1.5682	12	18.8184
39.5 44.5	42	1.6232	8	12.9850
			N=105	f log m =146.7490

$$GM = AL \frac{146.7490}{105} = AL\ (1.3976) = 24.98$$

Specific uses of G.M. : The Geometric Mean has certain specific uses some of them are :

It is used in the construction of index number.

It is also helpful in finding out the compound rates of change such as the rate of growth of population in a country.

It is suitable where the data are expressed in terms of rates, ratios and percentage.

Example 10. The gross national product of a company was Rs. 1,000 crores 10 years earlier. It is Rs. 2,000 crores now. Calculate the rate of growth in G.N.P.

Solution. In this case compound interest formula will be made applicable for computing the average annual per cent increase of growth. This formula is

$$P_n = P_0 (I + r)^n$$

where P_n = Principal sum (or any other variate) at the end of the period.

P_0 = Principal sum at the beginning of the period,

= rate of increase or

decrease. n = number of years.

It may be noted that the above formula can also be written in the following form:

$$r = \sqrt[n]{\frac{P_n}{P_0}}\ 1$$

Substituting the values given in the question in the formula, we have

$$r = \sqrt[10]{\frac{2000}{1000}} -1\ 1021\sqrt{} = AL\frac{\log_2}{10}--1 = AL \frac{0.30103}{10} - 1$$

AL (0.30103) -1 = 1.0718 - 1 = 0.718 or 7.18%.

Hence, the rate growth in G.N.P. = 7.18%

Example 11. The price of commodity increased by 5 per cent from 1948 to 1949, 8 per cent from 1949 to 1950 and 77 per cent from 1950 to 1951. The average increase from 1948 to 1951 is quoted at 26 per cent and not 30 per cent. Explain this statement and verify the arithmetic.

Solution. Taking P_n as the price at the end of the period P_0 as the price in beginning we can substitute the values of P_n and P_0 in the compound interest formula. Taking P_0 = 100 ; then P_n = 200.7

$P_n = P_0\ (I + r)^3$

$200.7 = 100\ (1 + r)_3$

$$\text{or } (1 + r) = {}^{3}\frac{200.7}{100}\ ,\ I + r = 3\sqrt{\frac{200.7}{100}}\ ,\ \sqrt{\frac{200.7}{100}}\ 1$$

= 1.260 = 0.260 = 26%

Thus increase is not average of (5 + 8 + 77). It is not 30 per cent. It is 26% as found out by G.M.

Weighted G.M. The weighted G.M. is calculated with the help of the following formula:

$$GM = n\sqrt{x_1 w_1\ \ x_2 w_2\ .\ .....\ x_n w_n}$$

$$= AL\ \frac{(\log x_1\ \ w_1)\ (x_2\ \ w_2)\ ....(\log x_n\ \ w_n)}{w_1\ \ w_2\ \ ....\ w_n}$$

$$= AL \frac{(\log x\ \ w)}{w}$$

Example 12. Find out weighted G.M. from the following data:

Group	Index Number	Weight
Food	352	48
Fuel	220	10
Cloth	230	8
House Rent	160	12
Miscellaneour	120	15

Solution :

Calculation of Weighted Geometric Mean

Group	Index Number	Weights	Log x	w log x
Food	352	48	2.5465	22.2320
Fuel	220	10	3.3424	23.4240
Cloth	230	8	2.3617	18.8936
House Rent	160	12	2.2041	26.4492
Miscellaneour	120	15	2.2788	34.1820
		93		225.1808

$$\text{G.M. (weighted)} = \frac{w \log x}{w}\quad \frac{22.1808}{93} = AL\ (2.4134)\ = \text{Rs. } 263.8\ .$$

1.9. Harmonic Mean

The harmonic mean is defined as the reciprocals of the values of the variable of the series. Symbolically,

$$\text{H.M.} = \frac{N}{\frac{1}{x_1}\ \ \frac{1}{x_2}\ \ \frac{1}{x_3}\ ....\ \ \frac{1}{x_n}}$$

$$\frac{N}{}$$

or $=\frac{1}{x_1}$

In the case of discrete series, the formula becomes

$$\text{H.M.}=\frac{N}{\frac{1}{x}}$$

and in the case of continuous series,

$$\text{H.M.}=\frac{N}{\frac{1}{m}}$$

It may be noted that none of the values of the variable should be zero.

Example 13, Calculate harmonic mean from the following data:

85, 70, 10, 75, 500, 8, 42, 250, 40 and 36.

Solution:

x	$\frac{1}{x}$
85	0.011765
70	0.014286
10	0.100000
75	0.013333
500	0.002000
8	0.125000
42	0.023810
250	0.004000
40	0.025000
36	0.027778
N=10	$\frac{1}{m}$ = 0.343372

$$\text{H.M.}=\frac{N}{\frac{1}{x}}=\frac{10}{0.343372}=28.9 \text{ approx.}$$

Example 14. From the following data compute the value of the harmonic mean:

y :	5	15	25	35	45
f :	5	15	10	15	5

x	f	$\frac{1}{x}$	$f . \frac{1}{x}$
5	5	0.200	1.000
15	15	0.067	1.005
25	10	0.040	0.400
35	15	0.029	0.435
45	5	0.022	0.110
	N=50		$f . \frac{1}{x} = 2.950$

Solution :

$$\frac{N}{f \quad \frac{1}{x_1}} \quad \frac{50}{2.950} = 17 \text{ approx ..}$$

Example 15: Calculate harmonic mean from the following frequency distribution:

x	f
0 10	5
10 20	15
20 30	10
30 40	15
40 50	5

Solution: First of all, we shall find out mid points of the various classes. They are 5, 15, 25, 35 and 45. Then we will calculate the H.M. by applying the following formula:

$$\text{H.M.} = \frac{N}{\frac{1}{m}}$$

The answer will be 17 (approx.).

Application of Harmonic Mean to Special cases: like Geometric means, the harmonic mean is also applicable to certain special types, of problems. Some of them are:

(i) If, in averaging time rates, distance is constants, then H.M. is to be calculated.

Example 16. A man travels 480 km. a day. One the first day he travels for 12 hours @ 48 km per hour and second day for 10 hours @ 48 km. per hour. Find out his average speed.

Solution. Here the harmonic mean and not the arithmetic mean, will be used.

$$\text{H.M.} = \frac{3}{\frac{1}{48} \quad \frac{1}{40} \quad \frac{1}{32}} \quad \frac{3}{\frac{37}{480}} \text{ 30km. per hour (approx.).}$$

The arithmetic mean would be $\frac{48+40+32}{3}$ = 40 km. per hour.

If, in averaging the price data, the prices are expressed as "quantity per rupee" then harmonic means is to be applied.

Example 17. A man purchase one kilo of cabbages from each of the four places at the rate of 20 kg., 16 kg., 12 kg., and 10 kg., per rupees respectively. On the average how many kilos of cabbages he has purchased per rupee Solution.

$$\text{H.M.} \quad \frac{N}{\frac{1}{x}} = \frac{4}{\frac{1}{20}\ \frac{1}{16}\ \frac{1}{12}\ \frac{1}{10}} \quad \frac{4}{\frac{12\ 15\ 20\ 24}{240}} \quad \frac{4}{\frac{71}{240}} \quad \frac{4\ 240}{71} \quad = 13.5 \text{ kg. per rupee.}$$

Self Assessment

Fill in the blanks:

The geometric mean of a series of n positive observations is defined as the ________ of their product.

The harmonic mean of n observations, none of which is ...., is defined as the reciprocal of the arithmetic mean of their reciprocals.

If all the observations of a variable are same, all the three measures of central tendency coincide, i.e., AM = GM = HM. Otherwise, we have ....

Self Check Exercise

What are the essentials of a good average?

Distinguish between Harmonic mean and Geometric Mean.

Define mean and mode.

1.10. Summary

Averages are the typical values around which other items of the distribution congregate. Averages are sometimes referred to as the measures of Central Tendency. They are useful for describing the distribution in concise manner, for comparative study of different distributions and for computing various other statistical measures such as dispersion, skewness and kurtosis and various other basic characteristics of a mass of data.

1.11 Glossary:

Average: An average is a single value within the range. of the data that is used to represent all the values in the series.

Arithmetic Mean: Arithmetic Mean is defined as the sum of observations divided by the number of observations.

Deciles : Deciles divide a distribution into 10 equal parts and there are, in all, 9 deciles denoted as D1, D2, ... D9 respectively

Geometric Mean : The geometric mean of a series of n positive observations is defined as the nth root of their product.

Harmonic Mean : The harmonic mean of n observations, none of which is zero, is defined as the reciprocal of the arithmetic mean of their reciprocals.

Measure of Central Value: Since an average is somewhere within the range of data it is sometimes called a measure of central value.

Median : Median of distribution is that value of the variate which divides it into two equal parts.

Mode: Mode is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed.

Partition Values: The values that divide a distribution into more than two equal parts are commonly known as partition values or fractiles.

Quartiles : The values of a variable that divide a distribution into four equal parts are called quartiles. Weighted Arithmetic Mean: Weights are assigned to different items depending upon their importance, i.e., more important items are assigned more weight.

1.12 Answers: Self Assessment

1. tables, frequency distributions
2. measure of central tendency
3. average
4. congregate
5. False
6. False :
7. True
8. True
9. True
10. (b) :
11. (a)
12. Mode
13. three times
14. nth root
15. zero
16. AM > GM > HM.
17. Refer to section 1.4
18. Refer to section 1.8 and 1.9
19. Refer to section 1.6 and 1.7

1.13 Terminal Questions

What is meant by central tendency? State important measures of central tendency.

Give relationship between A.M. GM and HM.

Calculate mean, median and mode from the following date :

Marks more than :	0	20	40	60	80	100	120
No. Of Students :	80	76	50	28	18	9	3

1.14 Suggessted Readings

Gupta, S.P. Statistical Methoods, Sultan Chand & Som, New Delhi.

Jain, T.R. and Aggarwal, S.C. Statistical Analysis, VK (India) Enterprises, New Delhi.

Gupta S.C. and Gupta Indra, Business Statistics, Himalaya Publishing House, Mumbai.

Hooda, R.P., Statistics for Business and Economics, Macmillan, New Delhi.

*****

Lesson – 2
Dispersion and Their Measures

Structure:

2.1 Learning Objectives:

After studying the lesson, you should be able to understand:

- **What is the Meaning and Definition of Dispersion.**
- **What are the different methods of studying Dispersion.**
- **Application of different measures of dispersion.**
- **Calculation of Mean Deviation, Standard deviation and Coefficient of variance.**

Introduction:

Measures of Central Tendency, Mean, Median, Mode etc., include the central position of the series. They indicate the general magnitude of the data but fail to reveal all the peculiarities and characteristics of the series. In other words, they fail to reveal the degree of the spread out or the extent of the variability in individual items of the distribution. This can be known by certain other measures, known as 'measures of Dispersion' or 'Measures' of Variation'.

This can be made clear with the help of the following example. Suppose there are three series, with the values as:

I	II	III
X	X	X
10	2	10
10	8	12
10	20	8
$\bar{X}\ \frac{X}{N},\frac{30}{3}$	$\bar{X}\ \frac{30}{3}$	$\bar{X}\ \frac{30}{3}$
X=30,	$\bar{X}$ 10	$\bar{X}$ 10
N=3, $\bar{X}$ 10		

In all the series, the value of simple Mean is 10. On the basis of this average we can say that the series are alike i.e., of the same type. But if we see the composition of the series, we find the following differences:

In the case of 1st series, the value are equal; but in 2nd and 3rd series, the values are unequal and do not follow any order.

The value of the deviation, item-wise, is quite different for the 1st, 2nd and 3rd series. But all these deviations cannot be noted or ascertained if the value of 'sample mean' is taken into consideration.

In these three series, it is quite possible that the value of Mean is 10; but the value of Median may be different from each other. This can be calculated as follows

I	II	III
10	2	8
10 Median	8 Median	10 Median
10	20	12

The value of 'Median' in 1st series is 10, in 2nd series = 8 and in 3rd series 10. Therefore, the value of Mean and Median are not identical.

Even though the average remains the same, the nature and extent of the distribution of the size of the items may vary. In other words, we can say that the structure of the frequency distributions may easily differ even if their means are identical.

Here it is a clear indication given by these points that for a complete and analytical study of the different series, reference must be made to the study of the deviation or scatter. In the absence of it, the study cannot be considered as complete.

Meaning and Definition of Dispersion:

The simplest meaning that can be given to the word 'dispersion' is a lack of uniformity in the sizes or quantities of the items of a group or series. This definition can be supplemented with the other definitions too.

'Reiglemen' has mentioned........ "Dispersion is the extent to which the magnitudes or qualities of the items differ; that is ; the degree of diversity."

The word dispersion may also be used to indicate the spread of the data.

In all these definition, we can find the basic property of dispersion that is "The value which indicates the extent to which all the value are dispersed about the central value in a particular distribution is called Dispersion or Variation or Scatter or Deviation."

Properties of a good measures of Dispersion:

There are certain pre-requisites that are essential for a good measure of dispersion:

It should be simple to understand.
It should be easy to complete.
It should be rigidly defined.
It should be based on each and every item of the distribution.
It should be capable of further algebraic treatment.
It should have sampling stability.
It should not be unduly affected by the
extreme items. Dispersion : Absolute or Relative:

The measures of dispersion can be either 'absolute' or 'relative'. Absolute measures of dispersion are expressed in the same units in which the original data are expressed. For example, if the series is expressed as "Marks' of the Student's in a particular subject; the absolute dispersion will provide the value in 'Marks'. The only difficulty in his connection is that if the two or more series are expressed in different units, the series cannot be compared on the basis of the dispersion.

'Relative' or 'Co-efficient' of dispersion is the ratio or the percentage of a measure of absolute dispersion to an appropriate average. The basic advantage of this measure is that two or more series can be compared with each other, even if they are expressed in different items/units.

Theoretically, "Absolute Measures' of dispersion are better. But from practical point of view, relative or co-efficient of dispersion is better and decibel to compare two or more series.

In the different mathematical formula to study dispersion, we can obtain absolute and relative dispersion. Nothing is specifically mentioned, we generally calculate relative or coefficient of dispersion.

Self Assessment :

Fill in the blanks:

1. The concept of .... related to the extent of scatter or variability in observations.

Measures of central tendency are known as the

Dispersion are also known as the ....

4. A good measure of dispersion should be ....defined.

5. A good measure of dispersion should be .... all the observations.

A good measure of dispersion should not be unduly affected by ....

State whether the following statements are true or false:

A measure of dispersion can be used to test the reliability of an average.

The extent of variability in two or more distributions can be compared by computing their respective dispersions. :

A distribution having lower value of dispersion is said to be more uniform or consistent.

State whether the following statements are true or false:

An absolute measure of dispersion is expressed in terms of the units of measurement of the variable.

A relative measure of dispersion, popularly known as coefficient of dispersion, is expressed as a pure number, independent of the units of measurement of the variable.

2.5. Methods of Dispersion:

The methods of studying dispersion can be divided in 'two' parts:

Methematical Methods: In these methods, we can study the 'degree' and 'extent' of dispersion. In these category, common measures of dispersion are:

(a) Range

(b) Quartile Deviation

(c) Mean Deviation or Average Deviation

(d) The standard deviation and co-efficient of variation.

II. Graphic Methods: Where we study only the 'Extent of Dispersion' whether it is more or less, but the 'degree' of dispersion is not possible. In this case, we take the help of the 'Lorenz-curve' or 'Cumulative percentage curve'.

(I) Methematical Methods:

2.5 (1a) Range: It is the simplest method of studying dispersion. It is defined as a value that gives the difference between the smallest value and the largest value of a series. It is clear in this case, that we do not take into account 'FREQUENCIES' of different groups, if any

Formula (Absolute Range) = L – S

(Co-efficient of Range) $\frac{L - S}{} = L + S$

Where : L = Largest value in a distribution

S = smallest value in a distribution.

Procedure for Calculating Range:

Take the 'largest' 'and 'smallest' values of a series but not frequencies.

Apply the formula of 'Absolute' and 'Relative' Range. For example, let us consider the following three series:

1. Raw Data : Marks of the students, in a class of 12 students, are given in Accounts as follows:

Marks in Account (out of 50)

12, 18, 20, 12, 16, 14, 30, 28, 12, 12 and 35.

In the said example, the maximum or the highest marks obtained by a candidates is '35' and the lower marks obtained by a candidate is '12'. Therefore, we can say that L = 35 and S = 12.

Absolute Range = L –S = 35 = 12 = 23 marks

Co-efficient of Range $\frac{L - S}{L + S}$ $\frac{35 - 12}{35 + 12}$ $\frac{23}{47}$ **49 (Approx.)**

II. Discrete Series :

Marks of the students in Accounts (out of 50) x		No. of students frequency (f)
Smallest	10	4
	12	10
	18	16
Largest	20	15
Total		45

Absolute Range = 20 - 10 = 10 Marks

Co-efficient of Range = $\frac{20 - 10}{20 + 10}$ $\frac{10}{30}$ **34 (approx.)**

III. Continuous Series

	X	frequencies
Smallest	10—15	4
Value S =10	15—20	10
	20—25	26
		Total = 48

Largest Value = (Upper limit of the last group)

= L=30

Absolute Range

= L-S = 30-10 = 20Marks.

Co-efficient of Range

$\frac{L - S}{L + S}$ $\frac{30 - 10}{30 + 10}$ $\frac{20}{40}$ **.5**

Merits of Range

It is the simplest method of studying dispersion.

It takes less time to compute the 'absolute' and 'relative' range. Demerits of Range

It does not take into account all the values of a series, i.e. it takes into account only the extreme items and middle figures are not given any importance. Therefore, 'Range' cannot tell us anything about the character of the distribution.

Range cannot be computed in the case of 'open ends' distribution i.e. a distribution where the lower limit of the first group and upper limit of the group or any of them is not even.

Uses of Range

The concept of range is useful in the case of quality control, to study the variation in the prices of the shares etc. 2.5 (I b) Quartile Deviation (Q.D.)

The concept of 'Quartile Deviation' does take into account only the values of the "upper quartile' (Q_3) and the 'Lower quartile' (Q_1). The Quartile Deviation may also be called as 'inter-quartile Range'. This is the good method when we are interested in knowing the range within which certain proportion of the items fall.

'Quartile Deviation' in its different forms can be obtained as:

Inter-Quartile range = $Q_3 - Q_1$

Semi-Quartile range $\frac{Q_3 \quad Q_1}{2}$

(c) Co-efficient of Q.D. = $\frac{Q_3 \quad Q_1}{Q_3 \, Q_1}$

Calculation of Inter-quartile Range, Semi-quartile Range and Co-efficient of Quartile Deviation in the case of Raw Data.

Suppose the values are :

X = 20, 12, 18, 25, 32, 10

In the case of quartile-deviation, it is necessary that the figure may be arranged in ascending or descending order before calculating the values Q_1 and Q_3.

Therefore, the arranged figures are : (in ascending order)

X=10, 12, 18, 20, 25, 32

No. of items = 6

Q_1 = The value of $\frac{(N+1)}{4}$ **th item,**

= The value of $\frac{(6+1)}{4}$ **th item.**

The value of 7/4 or 1.75th item.

The value of 1st item + .75 (value of 2nd item - value of 1st item)

10 + .75 (12 -10) =10 + .75 (2) = 10 + 1.50 = 1.50

Q_3 = The value of $\frac{3(N+1)}{4}$ **th item.**

= The value of $\frac{3(6+1)}{4}$ **th item.**

The value of 3 (7/4) th item.

The value of 21/4th item.

The value of 5.25th item.

The value of 5th item + .25

(The value of 6th item minus the value of 5th item)

= 25 + .25 (32 - 25) = 25 + .25 (7) = 25+ 1.75 or 26.75 = 26.75

Therefore,

(a) Inter-Quartile range $=Q_3 – Q_1$

$= 26.75\ 11.50 = 15.25$

(b) Semi-Quartile range $= \frac{Q_3\ \ Q_1}{2}$

$\frac{15.25}{2} = 7.625$

(c) Co-efficient of Q.D. $= \frac{Q_3\ Q_1}{Q_3\ \ Q_1}$

$\frac{26.75 - 11.50}{26.75 + 11.50}$

or $\frac{15.25}{38.25} = .36$ approx.

Calculation of Inter-quartile Range, Semi-quartile range and coefficient of quartile deviation in discrete series.

Suppose a series consists of the salaries (in Rupees) and Number of the Workers in a particular factory.

Salaries (in Rs.)	No. of Workers
60	4
100	20
120	21
140	16
160	9

In this problem, we are required to compute the values of Q_3 and $Q_{1.}$ They can be computed as follows :

Salaries (in Rs.) (x)	No. of workers (f)	Cumulative Frequencies (c.f.)	
60	4	4	
100	20	24	Q_1 less in this Cumulative Frequency
120	21	45	
140	16	61	
160	9	70	
	N = f = 70		

Calculation of Q_1

Q_1 is the size of $\frac{N+1}{4}$ th item = Size of $\frac{70+1}{4}$ th item

= Size of $\frac{71}{4}$ or 17. 75th item

17.75 lies in the cumulative frequency 24, which is corresponding to the value Rs. 100

Q_1 = Rs. 100

Calculation of Q_3

$$Q_3 = \text{Size of } \frac{N+1}{4} 3 \text{ th item}$$

$$Q_3 = \text{Size of} \frac{70+1}{4} 3 \text{ th item}$$

Size of 3(17.75)th item = Size of 53.25th item

5 3 .25 lies in the cumulative frequency 61, which is corresponding to Rs. 140.

Q_3 = Rs. 140

Inter-quartile Range

$Q_3 – Q_1$ = Rs. 140 – Rs. 100 = Rs. 40.

Semi-quartile Range

$$\frac{Q_3 \quad Q_1}{2} = \frac{\text{Rs. } 140 - \text{Rs. } 100}{2} = \text{Rs. } 20.$$

(c) Co-efficient of Q.D.

$$\frac{Q_3 \quad Q_1}{Q_3 \quad Q_1} = \frac{\text{Rs. } 140 - \text{Rs. } 100}{140+100}$$

$$= \frac{\text{Rs. } 40}{240} = .17 \text{ app.}$$

Calculation of Inter-quartile range, Semi-quartile range and Co-efficient of Quartile Deviation in case of Continuous Series.

Suppose the series is given as :

Salaries (in Rs.)	No. of Workers
10 20	4
20 30	6
30 40	10
40 50	5
	Total = 25

In this case, the value of Q_3 and Q_1 can be obtained as follows:

Salaries (in Rs.) (x)	No. of workers (f)	Cumulative Frequencies (c.f.)
10 20	4	4
20 30	6	10
30 40	10	20
40 50	5	25
	N =	

$$Q_1 \quad l \quad \frac{\frac{N}{4\text{-c.f.}}}{f} \quad i$$

where : l = Lower limits of Q_1 group

f = frequency of Q_1 group

i = magnitude of Q_1 group

c.f. = cumulative frequency of the group preceding

Q_1 group

N/4 is meant to find out Q_1 group.

Therefore, $\frac{N}{4}$ or $\frac{25}{4}$ or 6.25. It lies in the cumulative frequency 10, which is corresponding to class interval 20 — 30. Therefore, Q_1 group is 20—30.

$$Q_1 = 20 + \frac{6.2\text{-}4}{6} \quad 10$$

where : l = 20, f = 6, i = 10, n/4 = 6.25, c.f. = 4

or $Q_1 = 20 + \frac{2.25}{6} \quad 10 \quad 20\frac{2.25}{6} = 20 + 3.75 =$ Rs. 23.75

$$Q_3 = l + \frac{3N/4 \text{ - c.f. } i}{f}$$

$$3N/4 = \frac{3\ 25}{4} \quad \frac{75}{4} \quad 18.75$$

which lies in the cumulative frequency 20, which is corresponding to class interval 30 — 40.

Q_3 group is 30 — 40.

where l = 30, i = 10, 3n/4 = 18.75, c.f. = 10, f = 10

$$= 30 + \frac{18.75 \quad 10}{} \quad 10 \quad \text{Rs. } 38.75$$

$_3 10$

Therefore:

Inter-quartile range

= Q_3 Q_1 = Rs. 38.75 Rs. 23.75 = Rs. 15.00

Semi-quartile range

$$\frac{{}_3Q_1}{2} \quad \frac{15.00}{2} = \text{Rs.}7.50$$

(c) Coefficient of Q.D.

$$\frac{Q_3 Q_1}{Q_3 Q_1} \quad \frac{\text{Rs. } 38.75 \text{ - Rs.}23.75}{\text{Rs. } 38.75 + \text{Rs. } 23.75} = \frac{15}{62.50} = .24$$

Advantages of Quartile Deviation :

Some of the important advantages of this measure of dispersion are:

It is easy to calculate. We are required simple to find the values of Q_1 and Q_3 and then to apply the formula of absolute and coefficient of quartile deviation.

It has better results than range method. Because in the case of range, we take only the extreme values, than make dispersion erratic. In the case of quartile deviation we take into account middle 50% items.

The quartile deviation is not affected by the extreme items.

Disadvantages :

It is basically dependent on the values of the central items. If these values are ‘irregular’ and ‘abnormal’ the result is bound to be affected.

All the items of frequency distribution cannot be given equal weight or importance in finding the values of Q_1 and Q_3.

It does not take into account all the items of the series. Therefore, it is considered to be inaccurate measure of dispersion. So it should not be used, unless asked for.

Similarly, sometimes we calculate percentile range, say 90th and 10th as it gives slightly better measure of dispersion, in some cases. If we consider the calculation, then

(a) Absolute percentile range = P_{90} - P_{10}

(b) Coefficient of percentile range $\frac{P_{90} \quad P_{10}}{P_{90} \quad P_{10}}$

Suitability:

This method of calculating dispersion can be applied generally in the case of ‘open’ end series, where the importance of extreme values is not considered.

Self Assessment

Fill in the blanks:

The .... of a distribution is the difference between its two extreme observations.

A relative measure of range, also termed as the ....

State whether the following statements are true or false:

17041664034 Interquartile Range is an absolute measure of dispersion given by the difference between second quartile (Q3) and first quartile (Q1).

17041665034 Symbolically, Interquartile range = Q3 — Q1/2

2.5 (I c)III. Mean deviation or Average Deviation (M.D.)

‘Mean deviation’ is defined as a value, which is obtained by taking the average of the deviations of various items, from a measure of central tendency, Mean or Median or Mode, after ignoring negative signs.

Generally, the measure of central-tendency, from which the deviations are taken, is mentioned in the problem. If nothing is mentioned regarding the measures of central-tendency, in that case the deviations are taken from MEDIAN, because the sum of the deviations (after ignoring negative signs) is the minimum. Calculation in the case of Raw Data :

(a) Absolute Mean Deviation about Mean or Median or Mode $\frac{|d|}{N}$

where: N = Number of the observations.

d = deviations, that can be taken from Mean or Median or Mode

|d| = deviations after ignoring negative signs, i.e. all the ‘negative signs’ are converted into ‘positive signs’.

= Sum of

(b) Coefficient of M.D. =

$$\frac{\text{Mean Deviation about Mean/Median/Mode}}{\text{The value of Mean/Median/Mode}}$$

Steps:

%2289Calculate the value of Mean or Median or Mode.

%2289Take Deviations from the given measure of central tendency and they are shown as ‘d’.

%2289Ignore the negative signs of the deviation that can be shown as ‘| d |’ and add them up : so that the value | d | is obtained.

%2289Apply the formula to get Mean Deviation about

Mean or Median or Mode. Example. Suppose the values are given as 5, 5, 10, 15, 20. We want to calculate ‘Mean Deviation’ and Coefficient of Mean Deviation about Mean or Median or Mode.

Solution :

I. Mean Deviation (Absolute and Co-efficient) about Mean

X	Deviation from Mean ‘d’	Deviations after ignoring signs | d |
5	6	6
5	6	6
10	+ 1	1
15	+ 4	4
20	+9	9

X=55
N = 5
		| d| = 26

$$\overline{X} \quad \frac{X}{N}$$

where N=5, X=55

$$\frac{55}{5} \quad 11$$

Mean Deviation about Mean

$$\frac{|\,d\,|}{N} \quad \frac{26}{5} \quad 5.2$$

Co-efficient of Mean Deviation about mean

$$\frac{\text{Mean Deviation about Mean}}{\text{Value of Mean}}$$

$$\frac{5.2}{11} \; .47.$$

II. Mean Deviation (Absolute and Co-efficient) about Median :

	X	Deviation's from Median i.e. 13 'd'	Deviations after ignoring negative signs I d |
	5	5	5
	5	5	5
Median	$\boxed{10}$	0	0
	15	+5	5
	20	+10	10
	N= 5		| d| = 25

Mean Deviation about Mean $\frac{|d|}{N} = \frac{25}{5}$ = 5

Co-efficient of Mean Deviation about Median

$$= \frac{\text{M.D about median}}{\text{Value of Median}} \quad \frac{5}{10} \quad 5.$$

III. Mean Deviation (Absolute Co-efficient) about Mode :

X	Deviation's from Mode i.e. 5 'd'.	|d|
Mode[5]	0	0
5	0	0
10	+5	5
15	+10	10
20	+15	15
N = 5		| d| = 30

Mean Deviation about Mode.

$$\frac{|d|}{N} \quad \frac{30}{5} \quad 6.$$

Similarly, we can find out coefficient of Mean Deviation about Mode.
Coefficient of Mean Deviation

$$\frac{\text{M.D about median } 6}{\text{Value of Median } 5} \quad 1.2.$$

Calculation in the case of Discrete and Continuous series

Absolute M.D. about Mean or Median or Mode $\frac{f|d|}{N}$

where : N = No. of items
E = Sum of
f = frequency
| d | = deviation from Mean or Median or Mode after ignoring signs

Co-efficient of M.D. about Mean or Median or Mode

$$\frac{\text{M.D. about Mean or Median or Mode}}{\text{Value of Mean or Median or Mode}}$$

Example: Suppose we want to calculate Co-efficient of Mean Deviation about Mean from the following discrete series:

X	frequency
10	5
15	10
20	15
25	10
30	5

In this first of all, we shall be required to calculate the value of simple mean as follows :

Calculation of Mean (Simple)

X	f	fx	
10	5	50	$\overline{X}\ \frac{fX}{N}\ \frac{900}{45}\ 20$
15	10	150	
20	15	300	
25	10	250	
30	5	150	
	N=45	fx = 900	

Calculation of Co-efficient of Mean Deviation about Simple Mean

X	frequency f	Deviation for mean i.e. 20 'd'	Deviations ignoring negative signs | d |	Zf |d|
10	5	10	10	50
15	10	5	5	50
20	15	0	0	0
25	15	+ 5	5	50
30	5	+ 10	10	50
	N=45			f|d| = 200

co-efficient of Mean Deviation about Mean

$$\frac{\text{M.D about Mean}}{\text{Mean}}$$

M.D. about Mean = $\frac{f\,|\,d\,|}{N}\quad\frac{200}{45}$ **= 4.4 approx.**

Co-efficient of M.D. about Mean $\frac{4.4}{2}\ 2.2$

Suppose we want to calculate coefficient of Mean deviation about median from the following data :

Class-interval	Frequency
10—14	5
15—19	10
20—24	15
25—29	10
30—34	5
	N=45

In this case, we shall be required, first of all to calculate the value of Median but it is necessary that 'real limits' of the said class-intervals are to be obtained. This is possible only if, 5 is subtracted from the lower-limits and added to the upper limits of the given classes. Hence, the real limits are shown as follows:

9.5 — 14.5, 14.5 — 19.5, 19.5 — 24.5, 24.5 — 29.5 and 29.5 —34.5.

Calculation of Median

Class-interval	frequency 'f'	Cumulative frequency 'c.f'
9,5 — 14.5	5	5
14.5 — 19.5	10	15
19.5 — 24.5	15	30
24.5 — 29.5	10	40
29.5 — 34.5	5	45
	N =45	

$$\text{Median} = l \quad \frac{\frac{N}{2}\ \text{c.f.}}{f} \quad i$$

where: L = Lower limit of Median group
i = Magnitude of Median group
f = Frequency of Median group
c.f. = Cumulative frequency of the group preceding Median group

$\frac{N}{2}$ = is meant to find out Median group.

Therefore, we are required to calculate the value of $\frac{n}{2}$ i.e. $\frac{45}{2}$ 22.5

It lies in the cumulative frequency 30, which is corresponding to class interval 19.5 — 24.5. Median Group is 19.5 — 24.5.

Median group is = $19.5 + \frac{22.5 \quad 15}{15} 5 \quad 19.5 \qquad \frac{7.8}{15} 5$ = 19.5 + 2.5 = 19.5 +2.5 = 22.

Calculation of Co-efficient of Mean Deviation about Median

Class intervals	Frequency 'f'	Mid points X	Deviations about Median = 22 'd'	Deviations after igonoring signs | d |	f|d|
9.5 — 14.5	5	12	-10	10	50
14.5 — 19.5	10	17	-5	5	50
19,5 — 24.5	15	22	0	0	0
24.5 — 29.5	10	27	+5	5	50
29.5 — 34.5	5	32	+ 10	10	50
	N-45			f |d 1 = 200	

Coefficient of M.D. about Median

$$\frac{\text{~~M.D. about Median~~}}{\text{Median}}$$

$$\text{M.D. about Median} = \frac{f \mid d \mid}{N}$$

where N = 45, f |d | = 200

$$\text{M.D. about Median} = \frac{200}{45} \text{ 4.4. approx.}$$

$$\text{Coefficient of M.D. about Median} = \frac{4.4}{22} = .2$$

Advantages of Mean Deviations :

It takes into account all the items of a series. Therefore, it provides sufficiently representative results.

Since all the signs of the deviations are taken as positive, therefore, it simplifies calculations.

Mean Deviation may be calculated either by taking deviations from Mean or Median or Mode. It has got some specific advantages.

In this case, abnormality in the extreme items do not affect the value of Mean Deviation.

It is easy to calculate and to follow.

The measure obtained by this method may be used for making some comparisons.

Disadvantages of Mean Deviations :

It is illogical and mathematically not sound to convert all the negative signs into positive signs.

Since this method is not mathematically sound, so the results obtained by this series do not have any importance.

This method is not considered suitable for making comparisons either of the series itself or of the structure of the series.

Suitability:

This method is more effective in the reports presented to the general public or to groups not familiar with statistical methods.

2.5 (1d) Standard Deviations (Symbol =)

The concept of standard deviation, which is shown by Greek letter (read as sigma) is extremely useful in judging the representativeness of the mean. This concept was introduced by Karl-Pearson. The concept of standard deviation has a practical significance, because it is free from all those defects which exist in the case of range, quartiles deviation, mean deviation.

Standard deviation, is calculated as the square root of average of squared deviations from the actual mean. It is also called root mean square deviation. When the deviations are taken from the actual mean, but if the deviations are taken from the arbitrary mean, then standard deviation may not be called 'root mean square deviation'.

There is one significance point in this connection, that is, the square of the standard deviation i.e., $_2$ is called as ‘variance’ in statistics.

Calculation of standard deviation in the case of Raw Data:

In the case of Raw data, there are four ways of calculating standard deviation:

When the actual values are taken into account;

When the deviations are taken from actual mean;

When the deviations are taken from assumed mean ; and

When ‘step deviations’ are taken.

1. When the ‘Actual Values’ are taken into account:

$$\sqrt{\frac{\overline{X}}{N} \quad (\overline{X})^2}$$

or $\quad {}^2 \quad \frac{\overline{X}^2}{N} \quad (\overline{X})^2$

where N = No. of the items.

X = The value given in the series.

$\overline{X}$ = Mean of the value, that can be obtained with the help of $\frac{\overline{X}}{N}$.

Sometimes, this formula can also be presented as $\sqrt{\frac{\overline{X}_2}{\ } \quad \frac{X}{\ }^{N} \quad {}^2}$

where X = $\frac{\overline{X}}{N}$

Procedure to calculate:

Take the simple mean of the given values.

Take the squares of the actual values and add them up.

Apply the formula to get the value of standard deviation.

Example. Suppose the values are given as 2, 4, 6, 8, 10. We want to apply the formula

$$\sqrt{\frac{X_2}{N} \quad (X)^2}$$

In this case, we are required to calculate the values of N, $\overline{X}$, X_2. They are calculated as

X	X_2	$\sqrt{\frac{200}{5} \quad (6)^2}$
2	4	$\sqrt{44 \quad 36}$
4	16	$\sqrt{0\ 8}$
6	36	2.828
8	64	Variance $^2 \quad \sqrt{8}\ _2$
<u>10</u>	<u>100</u>	= 8

N=5, X_2=220, X =30,

$\underline{X}30_6$

$\frac{}{N5}$ $\_\_$

There are certain specific problems, where the method can be applied. If a different type of problem is given which is as follows:

Problems. In a distribution of 10 observations, the value of Mean and Standard Deviation are given as 20 and 8. By mistake, two values are taken as 2 and 6 instead of 4 and 8. Find out the value of correct mean and variance.

Solution. In the problem, we are given :

N=10, $\overline{X}$ = 20, = 8

Wrong values = 2 and 6

Correct values = 4 and 8

Required; to find out (a) correct Mean

(b).correct standard deviation

Calculation of correct Mean

$\overline{X}$ $\frac{X}{N}$ or $\overline{X}$ N X

X =10 20 =200

But this is wrong, because of the inclusion of wrong values in the series.

CorrectX =200-2-6+4+8=204

Correct Mean = $\frac{X}{N}$ $\frac{204}{10}$ 20.4

Calculation of correct variance :

$$\sqrt{\frac{X^2}{N}\ (\overline{X})^2}$$

$$\text{or}\quad {}^2\ \frac{X}{N}^2\ (\overline{X})^2$$

$$\text{or}\quad (8)^2\ \frac{X_2}{N}\ (20)^2$$

$$\text{or}\quad 64\ \frac{X}{10}^2\ 400$$

$$\text{or}\quad 64 + 400\ \frac{X^2}{10}\ \text{or } 464\ \frac{X^2}{10}$$

or X^2 = 4640.

But this is also wrong.

Correct X^2 = 4640-2_2-6_2+4_2+8_2

4640 - 4 -36+16 + 64

4640 - 40 + 80 = 4680

Correct $^2 = \frac{X^2}{N}$ - correct; $(\bar{X})_2$

$\underline{4680}$
- $(20.4)^2$ = 468 - 416.16 = 51.84

When the deviations are taken from Actual Mean :

$$=\sqrt{\frac{X^2}{N}}$$

Where, N = No. of items

X = Deviations from actual Mean

i.e. $(X - \bar{X})$

Steps :

Determine the deviations of different values from actual mean i.e., $X - \bar{X}$ that ate shown with the help of 'x'.

Square these deviations.

Add them.

Divide this sum by N.

Take the 'square root' to find the value of standard deviation.

Example. Suppose the values are given as 2,4,6, 8, 10. We want to apply the formula

$$=\sqrt{\frac{X^2}{N}}$$

Therefore,

X	x	x_2
2	26=4	$(4)_2$=16
4	46=2	$(2)_2$ =04
4	6 6=0	= 0
8	8 6=+2	$(2)_2$ =4
10	10 6=+4	$(4)_2$ = 16
N = 5		X_2=40

$\bar{X}$ = 6

$$= \sqrt{\frac{X^2}{N}} = \sqrt{\frac{40}{5}}\ 8\sqrt{2}.828$$

3. What the deviations are taken from Assumed Mean

$$\sqrt{\frac{X'^2}{N} \quad \frac{X^2}{N}}$$

where, N = No. of the items.

X = Deviations from Assumed Mean i.e. (X - A)

A = Assumed mean

Steps :

1. Take any value as assumed mean. This may be given in the series or may not be given in the series.
2. Take Deviations from the Assumed value i.e., (X - A), so that X is obtained in the series and add them up in order to get X .
3. Square the deviations obtained in point 2 and are shown as X_2 and add them up so that X_2 is obtained.
4. Apply the formula in order to get the value of standard deviation.

Example. Suppose the value are given as 2, 4, 6, 8, 10. We can obtain standard deviation with the help of the formula as:

	X	X’ = (X -A)	X'_2
	2	2=2 4	4
Assumed	$\boxed{4}$	0=4 4	0
Mean or A			
	6	2=6 4	4
	8	4=8 4	16
	10	6=10 4	36
	N = 5	X’=10	X'_2 = 60

$$\sqrt{\frac{X'^2}{N}} \quad \frac{X'^{\,2}}{N} \quad \sqrt{\frac{60}{5} \quad \frac{10^{\,2}}{5}} \quad 12\sqrt{4}$$

or $\sqrt{8}\sqrt{\ }$ 2.828

4. When step deviations are takes from the Assumed Mean :

$$\sqrt{\frac{d^{\,2}}{N} \quad \frac{d^{\,2}}{N}}$$

where : L = Common value

N = No. of items

d = Step deviations = $\frac{X'}{i}$

Steps :

Take the assumed mean from the given values.

Take Deviation from the assumed mean i.e., (X A), so that X is obtained.

Divide the deviations obtained in point 2 with a common value that is to say a value that can divide all the deviations from the assumed mean is called a ‘common value”.

This will give the value of d = $\frac{X'}{i}$ and add them up so that d is obtained.

The step deviations ‘d’ that are obtained in point 3 are squared and add them up so that d_2 is obtained.

Apply the formula to get the value of standard deviation.

Example. Suppose the values are 2, 4, 6, 8, 10. We can apply the formula as :

X	X =X-A	i =2
		d = $\frac{X}{i}$
			d2

2	2	1	1
A = 4	0	0	0
6	+ 2	1	1
8	+ 4	2	4
10	+ 6	3	9
N = 5		d = 5	d2 = 15

$$\sqrt{\frac{d^2}{N} \quad \frac{d}{N}^{2}} \; i$$

where, N = 5, i = 2, d = 5, $d^2 = 15$

$$\sqrt{\frac{15}{5} \quad \frac{5}{5}^{2}} \; 2 = 3 1 2 2 2 \sqrt{\quad} \quad \sqrt{\quad}$$

or = 1.414 = 2.828.

Note : We can see that the figures are same in all the four methods and the value of standard deviation is also same. Therefore, any of the four formulae can be applied to find the value of standard deviation. Of course, the suitability of a formula depends on the nature of items in a question.

Co-efficient of Standard-deviation = $\overline{X}$

In the given problem : = 2.828 and $\overline{X}$ = 6,

Therefore, coefficient of standard deviation $\frac{2.828}{6}$ .471

= Co-efficient of variation or (C. V.)

$$= \frac{\quad}{\overline{X}} \; 100 \qquad \frac{2.828}{6} \; 100 \;\; 47.1\%$$

Generally, coefficient of variation can be obtained to compare two or more series on the basis of coefficient of variation. If coefficient of variation (C.V.) is more in one series as compared to the other there will be more variations in that series, less stability or consistency in the composition. If it is less, as compared to other series, it will be more stable, or consistency in the values. Moreover, that series is always better where coefficient of variation is less or coefficient of standard deviation is less.

Example. Suppose we want to compare two firms where the salaries of the employees are given as follows:

	Company A	Company B
1. No. of workers	100	100
2. Mean salary (in Rs.)	100	80
3. Standard-deviation (in Rs.)	40	45

In this case, we can compare them either with the help of coefficient of standard deviation or coefficient of variation. If we take coefficient of variation, then we are required to apply the formula :

C.V. = = $\overline{X}$ 100

Firm A	Firm B
C.V. $\frac{40}{100}$ 100	C.V. $\frac{45}{80}$ 100
$\overline{X}$ = 100, = 40	$\overline{X}$ 80, = 45
C.V. = 40%	C.V. $\frac{45}{80}$ 100 = 56.25%

We can easily conclude that 'coefficient of variation' is less in Firm A as compared to Firm B. Therefore, Firm A is better.

Calculation in Discrete and Coefficient Series :

The formula for calculating standard deviation in continuous series and discrete series are the same. The only difference, theoretically can be observed that in discrete series, values and frequencies are given; whereas in continuous series, class-intervals and frequencies are given. When the mid-points of these class-intervals are obtained, a continuous series takes the form of discrete series. The symbol 'X' denotes values in discrete series and mid-points in the case of continuous series,

Calculations in 'discrete' and 'continuous' series, when the deviations are taken from Actual Mean :

$$\sqrt{\frac{fX^2}{N}}$$

where N = Number of items = f

= frequency corresponding to different values in discrete series and 'class-intervals' in continuous series.

= Deviations from actual mean i.e. (X - $\overline{X}$)

X = 'Values' in discrete series and 'mid-points' in continuous series.

Steps:

Take the value of the arithmetic mean by applying any formula for calculating simple mean.

Take deviations from the 'Simple Mean' obtained in point 1, so that these deviations can be shown with the help of x.

Square the deviations obtained in point 2, so that the values of x_2 are obtained.

Multiply the frequencies of the different class-intervals with x_2 of different values obtained in point 3 so that 'fx_2' is obtained corresponding to different groups and add them up, so that f is obtained.

Apply the formula to get the value of standard deviation.

Note : If we want to calculate variance then we can take $\frac{fX^2}{N}$

Example. Suppose we are given the series as follows, where we want to calculate standard deviations by taking deviations from actual mean.

Class Intervals	Frequency 'f'	Mid-points 'x'	fx
10—14	5	12	60
15—19	10	17	170
20—24	15	22	330
25—29	10	27	270
30—34	5	32	160
	Total N = 45	fx = 990	

$$- \frac{fX}{N}$$

where : N =45,   fx = 990

$$X - \frac{900}{45} \quad 22$$

Calculation of Standard Deviation

Class Intervals	f	Mid points	Deviations from actual mean = 22 X	X^2	fx_2
10—14	5	12	-10	100	500
15—19	10	17	-5	25	250
20—24	15	22	0	0	0
25—29	10	27	+ 5	25	250
30—34	5	32	+ 10	100	500
	N=45			fx_2 = 1500	

$$\sqrt{\frac{fX^2}{N}}$$ where : N = 45, fx_2 = 1500

$$\sqrt{\frac{1500}{45}} \quad \sqrt{33.33}\ 5.77$$ approx.

Where the deviations are taken from Assumed Mean :

In some cases, the value of simple mean be in fraction, ‘there it becomes complicated to take deviations and then square them. Therefore, in that case deviations can be taken from the assumed mean.

$$\sqrt{\frac{fx'^2}{N} \qquad \frac{fx'^2}{N}}$$

where N = Number of items

x = Deviation from Assumed Mean i.e. (X - A)

f = frequency of the different groups

A = Assumed Mean

X = Values of mid points.

Steps

Take the assumed mean from the given values of mid-points.

Take Deviations from the assumed mean so that the value of X is obtained.

Square the deviations obtained in point 2 above.

Multiply ‘f’ with X of different groups so that the product fX is obtained corresponding to different groups and add them up so that fX is obtained.

Multiply T with X_2 of different groups so that the product fX is obtained for different groups and add them up, so that fX_2 is obtained.

Apply the formula to get the value of standard deviation.

Example. Suppose we are given the series and we want to apply this formula, then we are required to calculated the values of N, fX_2 and fX .

Class Intervals	Frequency f	Mid points X	Deviations from assumed Mean = 17 X	X_2	fX	fX_2
10—14	5	12	-5	25	-25	125
15—19	10	17	0	0	0	0
20—24	15	22	+ 5	25	75	375
25—29	10	27	+ 10	100	100	1000
30—34	5	32	+ 15	225	75	1125
	N=45			fX = 225,fX_2 = 2625		

$$\sqrt{\frac{fx'^{2}}{N}\qquad\frac{fx'^{2}}{N}}$$

where N = 45, fX_2 = 2625, fX = 225

$$\sqrt{\frac{2625fx'^{2}}{45}\qquad\frac{225_{2}}{45}}\quad 58.33\sqrt{2533.33}\;5.77\sqrt{\text{approx.}}$$

When ‘the step deviations’ are taken from the Assumed Mean :

$$\sqrt{\frac{fd'^{2}}{N}\qquad\frac{fd}{N}^{\;2}}\qquad i$$

where i = common value

N = Number of items = f

f = frequencies corresponding to the different groups.

d = Step-deviations = $\frac{X'}{i}$

X = (X A) : Deviations from assumed mean.

Steps :

1. Take Deviations from the assumed mean, so that they are symbolised as X .

Take common value from these deviations that are taken from assumed mean and the common value can be shown with the help of ‘i’.

Divide the deviations that are taken in point 1 by the common value ‘i’ so that the step deviations are obtained for different groups and they are denoted as ‘d’.

Square these step deviations so that ‘d$_2$’ is obtained for different groups.

Multiply ‘f’ and ‘d’ of different groups so that ‘fd’ is obtained and add them up so that fd is obtained.

Multiply’ f’ with ‘d$_2$' of the different groups so that ‘ fd’ is obtained for different groups and add them up so that fd$_2$ is obtained.

Apply the formula to get standard deviation.

Example. Suppose we are given the series and we want to calculate ‘standard deviation’ with the help of step deviation method. According to the given formula, we are required to calculate the value of i, N, fd and fd$_2$.

Class Intervals	f	Mid points	Deviations from assumed Mean i.e., 22	i = 5 $d = \frac{X'}{i}$ x'	d^2	fd	fd^2
10—14	5	12	–10	–2	4	–10	20
15—19	10	17	– 5	1	1	–10	10
20—24	15	22	0	0	0	0	0
25—29	10	27	+ 5	+ 1	1	10	10
30—34	5	32	+ 10	+ 2	4	10	20
	N = 45					fd = 0	fd^2 = 60

$$\sqrt{\frac{fd^2}{N} \qquad \frac{fd}{N}^{2}}\; i$$ where N= 45, i = 5, fd = 0, fd ^{2} = 60

$$\sqrt{\frac{60}{45} \quad \frac{0}{45}^{2}}\; 5 = \sqrt{\frac{4}{3}}\; 5 \quad \sqrt{1.33} \quad 5$$

or = 1.154 5 = 5.77 approx.

The combined standard deviation can be obtained for two or more series with the help of the following formula:

$$_{12} \quad \sqrt{\frac{N_1 \,_1^2 \; N_2 \,_2^2 \; (\bar{X}_1 \; \bar{X}_{12})^2 \qquad N_2 (\bar{X}_2 \; \bar{X}_{12})^2}{N_1+N_2}}$$

N_1 and N_2 are the number of items of two series

$_1$ and X_2 = Means of 1st and 2nd series

$_{12}$ = combined Mean of two series.

$_{12}$ = combined standard deviation of two series $_1$

and $_{12}$ = variations of 1st and 2nd series.

For example we want to find out combined standard deviation of two series, where the information is given as follows:

	1st Series	2nd Series
No. of items	10	15
Means of the series	15	20
Standard deviation	4	5

Since the two series are involved, therefore combined standard deviation is:

$$_{12} \quad \sqrt{\frac{N_1 \,_1^2 \; N_2 \,_2^2 \; N_1 (\bar{X}_1 \; \bar{X}_{12})^2 \qquad N_2 (\bar{X}_2 \, \bar{X}_{12})^2}{N_1+N_2}}$$

where

$N_1 = 10$, $N_2 = 15$,

$\bar{X}_1 = 15$, $\bar{X}_2 = 20$,

$_1 = 4$, $_2 = 5$,

$$\bar{X}_{12} \quad \frac{\bar{X}_1 N_1 \quad \bar{X}_2 N_2}{N_1 + N_2}$$

$$\text{or } \bar{X}_{12} \quad \frac{(15\ \ 10)\ \ (20\ \ 15)}{10+15} \quad \frac{150\ \ 300}{25} \quad \frac{450}{25}\,18$$

By applying the formula of combined standard deviations, we get:

$$_{12} \quad \sqrt{\frac{10(4)^2\ \ (15(5)^2\ \ \ 10(15\ \ 18)^2\,15(20\ \ 18)^2}{10+15}}$$

$$\sqrt{\frac{(10\ \ 16)\ \ (15\ \ 25)\ \ (10\ \ 9)\ \ (15\ \ 4)}{25}} \quad \sqrt{\frac{160\ \ 375\ \ 90\ \ 60}{25}}\,27\sqrt{.4\ \ 5}.2$$

Merits of Standard Deviation :

The standard deviation is the measure of dispersion, because it takes into account all the items and is capable of further algebric treatment.

It is possible to calculate standard deviation for two or more series.

For comparison, on account of variations, this measures is more suitable.

It can be used as a tool for statistical analysis.

Demerits :

It is difficult to compute.

It gives more weight to extreme items and less to those that are near the mean. It is on account of the fact that the squares of the deviations which are big in size would be proportionately greater than the square of those deviations which are comparatively small.

2.5 (2) Graphic Methods of Studying Dispersion :

The concept of Lorenz-curve was devised by Max-o-Lorenz, as a graphic method of studying dispersion. It is also called a cumulative percentage curve, in which the percentage of the items is combined with the percentage of other things as wealth, profits, etc.

While drawing Lorenz-curve, the following procedure is adopted:

The size of the items and the frequencies are converted into percentages.

Cumulative percentage are obtained for the values and frequencies.

On OX axis we take into account 'cumulative frequencies' into percentage and divide them in equal parts. On OX axis we start from 100% and move to 0%.

On OY axis, we take account 'cumulative values in percentage' and divide the axis in equal parts in such a manner so that the different parts on both the axis are equal. On OY axis we start from 0 to 100.

Join 0% on OX axis and 100% on OY axes so that a straight line is obtained which is called 'Line of Equal Distribution'.

We plot points of different series, on the basis of cumulative values in percentage and cumulative frequencies in percentage and join these points with freehand curve, so that 'Lorenz curve' is obtained for different series.

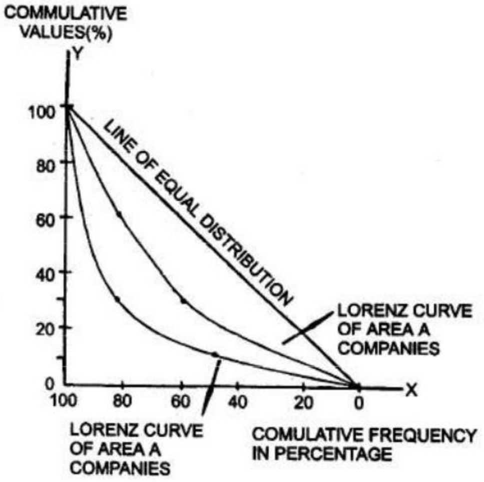
The' Lorenz-curves' are compared with the' Line of Equal Distribution' and the distance between them will decide about dispersion. More the distance, more the dispersion in the series will be.

Example. In the table given below is the number of companies belonging of two areas A and B. According to the amount of profits earned by them. Draw in the same diagram their Lorenz curves and interpret them.

	No. of Companies	
Profit earned (in Rs. lakhs)	Area A	Area B
10	4	24
20	10	18
30	6	5
40	5	3

In this problem, the solution can be given as follows:

X Profits earned	f_A f_B Number of Companies		Percentage		Cumulative Percentage				
(in Rs. lakhs)	Area A	Area B	X %	f_A %	f_B %	cx %	cf_A %	cf_B %
10	4	24	10	16	48	10	16	48
20	10	18	20	40	36	30	56	84
30	6	5	30	24	10	60	80	94
40	5	3	40	20	06	100	100	100
Total : 100	25	30	100	100	100			



In this diagram we see that the Lorenz Curve for 'Area A' in companies is away from the line of equal distribution as compared to Lorenz curve for Area B. Therefore, we conclude that there are more variations in Area B as compared to Area A.

In this graphic method there is only one limitation that is the value of dispersion cannot be obtained, but can simply get extent of dispersion i.e., whether it is more or less.

(C) Self Assessment :

Multiple Choice Questions:

16. .... is a Measure of dispersion based on all the observations.

Mean (b) Mean deviation (c) Quartiles (d) Standard deviation

Mean deviation is defined as the .... of the absolute deviations of observations from a central value like mean, median or mode.

Mean (b) Arithmetic mean (c) Geometric mean (d) Harmonic mean

State whether the following statements are true or false:

The concept of standard deviation was introduced by Karl Pearson in 1897.

A relative measure of dispersion, based on standard deviation is known as coefficient of Standard deviation.

(D) Self Check exercises :

Discuss the properties of a good measure of dispersion.

What do you understand by Mean deviation?

REVISION PROBLEMS

Example 1.

Compute (a) Inter-quartile range

(b) Semi-quartile range

(c) Co-efficient of quartile deviation from the following data:

Farm Size (acres)	No. of farm
0—40	394
41—80	461
81 — 120	391
121 — 160	334
161 — 200	169
201 — 240	113
241 — and over	148

Solution :

In this case, the real limits of the class intervals can be obtained by substracting 0.5 from the lower limits of the class intervals and adding .5 to the upper limits of the different class intervals. This adjustment is necessary while calculating Median and quartile in the series.

Calculation of Lower Quartile (Q_1) and Upper Quartile (Q_3)

Farm Size (acres)	No. of farm f	Cumulative frequency c.f.
0— 40	394	394
41— 80	461	855
81 — 120	391	1246
121 — 160	334	1580
161 —200	169	1749
201 —240	113	1862
241 — and over	148	2010
	N-2010	

$$Q_1 \quad L \quad \frac{\frac{N}{4}\ \text{c.f.}}{f} \quad i$$

$$\frac{N}{4} \quad \frac{2010}{4} \quad 1502.5 \text{ th item.}$$

It lies in the cumulative frequency of the group 41 — 80, where the real class intervals arc 40.5 — 80.5.

L = 40.5, f = 461, i =

40 c.f. = 394, n/4 = 502.5

$$Q_1\ 40.5 \quad \frac{502.5\ \ 394}{461} \quad 40 \qquad = 40.5 + 9.4 = 49.4 \text{ acres}$$

$$Q_3 \quad L \quad \frac{\frac{3N}{4}\ \text{c.f.}}{f} \quad i$$

$\frac{3N}{4}$ $\frac{3\ 2010}{4}$ 1507.5 th item

It lies in the cumulative frequency of the group 121 — 160, where the real limits of the class intervals are 120.5 — 160.5.

Therefore : L = 120.5, i = 40, f = 334, 3n/4 = 1507.5, c.f. = 1246

Q_3 120.5 $\frac{1507.5\ \ 1246}{3344}$ 40 = 120.5 + 31.3 = 151.8 acres

Inter-quartile range

%4 u $Q_3 - Q_1$ = 151.8 – 49.9

= 101.9 acres Semi-quartile range

$\frac{Q_3 Q_1}{2}$ $\frac{151.8\ \ 49.9}{2}$ = 50.95 acres

Co-efficient of quartile deviation

$\frac{Q_3 Q_1}{Q_3\ Q_1}$ $\frac{151.8\ \ 49.9}{151.8\ \ 49.9}$ $\frac{101.9}{201.7}$ = .5 approx.

Example 2. Calculate Mean and Co-efficient of Mean Deviation about Mean from the following data:

Marks less than	No, of students
10	4
20	10
30	20
40	40
50	50
60	56
70	60

Solution :

In this question, we are given 'less than type series' alongwiththe 'cumulative frequencies'. Therefore, we are required to find out class intervals and frequencies for calculating Mean and co-efficient of Mean Deviation about Mean.

	Marks of Students	No. of points	Mid Point from assumed Mean N = 35	Deviations i=10	Step Dev. from Mean = 35 (ignoring	Deviations
	f	x	x_1	d =	signs) | d | fd	f |d|
0 —10	4	5	–30	–3	3 –12	12
10—20	6	15	–20	–2	2 –12	12
20—30	10	25	–10	–1	1 –10	10
30—40	20	35	0	0	0 0	0
40—50	10	45	+ 10	+ 1	1 +10	10
50—60	6	55	+ 20	+ 2	2 +12	12
60—70	4	65	+ 30	+ 3	3 +12	12
	N=60				fd = 0	f |d | = 68

$\bar{}$%4 u $\frac{\quad}{N}$A^{fd} i where : N = 60, A = 35, i = 10, fd = 0

$X\ \overline{55}\ \frac{0}{60}\ 10\ 35$

M.D. about Mean $\frac{f\,|d|}{N}\ i \quad \frac{68}{60}\,10 = 11.33$

Co-efficient of M.D. about Mean $= \frac{\text{M.D about Mean}}{\text{Mean}}$

$1\underline{1.33}_{35}$ = .334 approx.

Example 3. An analysis of the monthly wages paid to the workers in two firms A and B belonging to the same industry, gives the following results:

	Firm A	Firm B
Number of wage earners	200	100
Average monthly wages (in Rs.)	50	45
Variance of the distribution of wages	100	121

%4 u Which firm, A or B, pays out the larger amount as monthly wages ?

%4 u In which firm, A or B, is there greater variability in individual wages ?

%4 u Find out the average monthly wages and standard deviation of the wages of all the workers in the two firms.

Solution : (a) In this questions, we are required to calculate total wages in Firm A and Firm B. Total wages can be obtained by multiplying Number of Workers with the average monthly wages, Therefore, in Firm A, total wages are 200 50 = Rs. 10,000 and in the firm B, the total wages are 100 45 = Rs. 4,500. Therefore, in firm A, more wages are being paid.

(b) In order to find out greater variability, we are required to calculate co-efficient or variation can be obtained with the help of $\frac{=}{X}\ 100$.

Therefore, calculation of co-efficient of variation is shown below:

	Firm A	Firm B
Given	$\overline{X} = 20$	$\overline{X} = 45$
	$^2 = 100$	$^2 = 121$
or	$= \sqrt{100} = 10$	$= \sqrt{121} = 11$

Co-efficient of variation

Firm A	Firm B
$= \overline{\overline{X}}\ 100$	C.V. $= \frac{11}{45}\ 100$
$= \frac{10}{50}\ 100$	$= \frac{1100}{45}$
= 20%	= 24.4%

It is clear from here that co-efficient of variation is more in Firm B as compared to Firm A. Therefore, there is greater variability in Firm B.

$$\text{(c) } \bar{X}_{12} \quad \frac{\bar{X}_1 N_1 \quad \bar{X}_2 N_2}{N_1 + N_2}$$

$$\frac{\text{Rs .}(200 \quad 50) \quad (100 \quad 45)}{200+100} \quad \frac{10,000 \quad 4,500}{300} \quad \frac{14,500}{300} = \text{Rs. } 48.33$$

$$_{12} \sqrt{\frac{\%4 \quad u \quad {}_1 \, 1^2 \quad N_2 \, \bar{} \, 2^2 \, \bar{} \, N_1(X_1 \quad X_{12})^2 \quad N_2(\bar{X}_2 \bar{X}_{12})^2}{N_1 N_2}}$$

$$\sqrt{\frac{(200 \quad 100) \quad (100 \quad 121) \quad 200(50 \quad 48.33)^2 \quad 100(45 \quad 48.33)^2}{200+100}}$$

$$\sqrt{\frac{(20,000 \quad 12,100 \quad (200)(1.67)^2 \quad 100(3.33)^2}{300}}$$

$$\sqrt{\frac{(20,000 \quad 12,100 \quad 557.8 \quad 1108.89)}{300}}$$

$$\sqrt{\frac{33,766.69}{300}} \quad 11\sqrt{2.5556} \quad 10.61 \text{ approx.}$$

Example 4 : (a) Find out the co-efficient of variation of a series for which the following results are given:

N = 50, x = 25, X_2 = 500

where : X = deviation from the assumed average 5.

For a frequency distribution of marks, in statistics of 100 candidates (grouped in class intervals of 0 — 10, 10 — 20) the mean and standard deviation were found to be 45 and 20. Later it was discovered that the score 54 was misread as 64 in obtaining frequency distribution. Find out the correct Mean and correct a standard deviation of the frequency distribution.

Can coefficient of variation be greater than 100%? If so, when?

Solution, (a) We want to calculate, co-efficient of variation, which is = X 100 .

Therefore, we are required to calculate the values of mean and standard deviation.

Calculation of simple mean :

$$X \quad A \quad \frac{x'}{N}$$

where: A = 5, N = 50, x = 25

$$X = \bar{5} + 50^{25} \text{—} \quad 5.5$$

Calculation of Standard Deviation :

$$\sqrt{\frac{x'^{2}}{N} \qquad \frac{x'^{2}}{N}} = \sqrt{\frac{500}{50} \qquad \frac{25}{50}^{2}} \qquad \sqrt{10\ .25} \quad \sqrt{9.75}\ \ 3.179$$

Calculation of co-efficient of variation

$$c.v = \frac{}{\overline{X}}\ 100 \quad = \frac{3.179}{5.5} \quad 100 \quad 57.8\%.$$

(b) Given $\overline{X}$ = 45, = 20, N = 100

Wrong value = 64, correct value = 54

Since this is the case of a continuous series, therefore, we will apply the formulae for Mean and Standard Deviation that are applicable in continuous series.

Calculation of correct mean :

$$\overline{X} \quad \frac{fX}{N} \text{ or } N\ \overline{X} = fx$$

By substituting the values, we get 100 45 = fx = 4500

Correct fx = 4500 - 64 + 54 = 4490

$$\text{Correct} \quad \overline{X} = \frac{\text{Correct } fx}{N} \quad \frac{4490}{100} 44.69$$

Calculation of correct σ :

$$\sqrt{\frac{fx^2}{N} \qquad \overline{X}_2}$$

$$\text{or} \qquad \frac{fx^2}{N} \qquad \overline{X}_2$$

where : = 20, N = 100, $\overline{X}$ = 45

$$(20)^2 \quad \frac{fx^2}{100} \quad (45)^2$$

$$\text{or } 400 = \frac{fx^2}{100}\ 2025^2$$

$$\text{or } 400 + 2025 = \frac{fx^2}{100}$$

or 2425 100 = fx^2 = 242500

Correct fx^2 = 242500 $(64)^2$ + $(54)^2$ = 242500 4096 +2916

= 242500 1180 = 241320

$$\text{Correct} \quad = \sqrt{\frac{\text{Correct } fx^2}{N}} \quad (49.9)^2 \qquad \sqrt{\frac{24132090}{100}\ (49.9)^2}$$

$$\sqrt{\sqrt{2413.20 \quad 2016.01}} \quad \sqrt{\sqrt{397.19}} \quad 39.9 \text{ approx.}$$

(c) The formulae for the computation of co-efficient of variation is $\frac{100}{\overline{X}}$

Hence, co-efficient of variation can be greater than 100% only when the value of standard deviation is greater than the value of Mean.

This will happen when data contains a large number of small items and few items are quite small. In this case the value of simple mean will be pulled down whereas the value of standard deviation will go up.

Similarly, if the negative items are there in the series, the value of mean will come down and the value of standard deviation shall not be affected, because of squaring the deviations.

Example 5. Calculate standard deviation from the following data :

Class intervals	frequencies
— 30 to — 20	5
— 20 to — 10	10
— 10 to 0	15
0 to 10	10
10 to 20	5
	N=45

Solution

Calculation of Standard Deviation

Class	Frequency	Mid	Deviations Step		Deviations		
Intervals		points	from assumed				
Mean = - 5			when i = 10				
	f	X	x	d = $\frac{X'}{i}$	d^2	fd	fd_2
— 30 to — 20	5	25	20	2	4	10	20
— 20 to — 10	10	15	10	1	11	10	10
— 10 to 0	15	5	0	0	0	0	0
0 to 10	10	5	+10	1	1	10	10
10 to 20	5	15	+ 20	2	4	10	20
	N=45					fd = 0	fd2=60

$$\sqrt{\frac{fd^2}{N} \quad \frac{fd}{N}^{\ 2}} \quad i$$

where : N = 45, i = 10, fd = 0, fd_2 = 60.

$$\sqrt{\frac{60}{45} \quad \frac{0}{45}^{\ 2}} \qquad 10\sqrt{\frac{60}{45}} \quad 10$$

$$\sqrt{\frac{4}{3}} \qquad 10\sqrt{1.33} \quad 10 \quad 1.153 \quad 10 \quad 11.53$$

2.6 Summary

Dispersion is the measure of the variation of the items. Measures of dispersion enable us to find out the reliability of an average, to control the variation of the data from the central value, to compare two or more set of data regarding their variability and for computing other statistical measures which are used extensively like correlation analysis, regression analysis the testing hypotheses, analysis of variance etc.

2.7 Glossary:

Dispersion: Dispersion is the measure of extent to which individual items vary.

Averages of second order : The measures which express the spread of observations in terms of the average of deviations of observations from some central value are termed as the averages of second order, e.g., mean deviation, standard deviation.. etc.

Interquartile Range: Interquartile Range is an absolute measure of dispersion given by the difference between third quartile (Q3) and first quartile (Q1) Symbolically, Interquartile range = —6Q_3 - Q_1 .

Quartile deviation or semi-interquartile range: Half of the interquartile range is called the quartile deviation or semi-interquartile range.

Range: The range of a distribution is the difference between its two extreme observations, i.e., the difference between the largest and smallest observations. Symbolically, R = L - S where R denotes range, L and S denote largest and smallest observations. |

Standard deviation: It is the square root of average of squared deviation taken from actual mean. This measure of dispersion is known as standard deviation or root-mean square deviation. Coefficient of standard deviation: A relative measure of dispersion, based on standard deviation is known as coefficient of standard deviation.

Variance: Square of standard deviation is known as variance :

2.8 Answers : Self Assessment

1. dispersion 2. averages of first order 3. averages of second order
4. True 5. True 6. True
7. rigidly 8.basedon 9. extreme observations
10. True 11. True 12. Range
13. Coefficient of range 14. False 15. False
16. (b) 17. (b) 18. False
19. True 20. Refer to section 2.3 and 2.4 7 21.Refer to section 2.5(IC)

2.9. Terminal Questions :

Q.1: Find the mean and variance of a distribution in which the value are 1, 2, 3, ..., n. and frequency of each is unity.

Ans: Mean = $\frac{1}{2}$ (n + 1), Variance = $\frac{n^2\ \ 1}{12}$.

Q.2: Calculate the S.D. and coefficient from the following table by step deviation method:

Class :	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Frequency:	3	61	132	153	140	51	2

Ans: S.D. = 11.84, coeff of S.D. = 0.216

Q.3. The monthly income of five labourers in rupees as 30, 40, 45, 50, 55. Find the deviation from the median. Ans. 7

Q.4. Find the quartile deviation from the following data:

Size:	1	2	3	4	5	6	7	8	9	10
Frequency	3	2	5	7	9	5	8	10	2	1

Ans. 1.725.

Q5. What is meant by dispersion? What purpose does a measure of dispersion serve?

2.10 Suggested Readings:

Heinz Kohler, Statistics for Business and Economics, Harper Collins.

Hooda, R.P. Statistics for Business and Economics, Macmillan, New Delhi.

Levin, Richard I and David S. Rubin : Statistics for management, Prentice Hall, Delhi.

Gupta S.C. & Gupta Indra, Business Statistics, Himalaya Publishing House, Mumbai.

*****

Lesson-3

Measures of Skewness and Kurtosis

Structure:-

3.1 Learning Objectives:

After studying the Lesson, you should be able to understand:

%4 u The meaning and Nature of Skewness

%4 u Different Methods of Measuring Skewness

%4 u Measure of Kurtosis.

3.2. Introduction:

Measures of Skewness and Kurtosis, like measures of central tendency and dispersion, study the characteristic of a frequency distribution. 'Average' tell us about the central value of the distribution and measures of dispersion tell us about the concentration of the items round a central value. These measures do not reveal whether the dispersal of value on either side of an average is symmetrical or not. If observations are arranged in a symmetrical order round a measure of central tendency, we get a 'symmetrical distribution', otherwise, it may be arranged in an asymmetrical order which gives asymmetrical distribution. Thus, Skewness is a measure that studies the degree of departure from symmetry.

A symmetrical distribution, when presented on the graph gives a 'symmetrical curve', where the value of Mean, Median and Mode are exactly equal. On the other hand, in an asymmetrical distribution, the values of Mean, Median and Mode are not equal.

When two ore more symmetrical distribution are compared, the difference in them are studied with KURTOSIS'. On the other hand, when two or more asymmetrical distributions are compared, it will give different degrees of 'SKEWNESS'. These measures are mutually exclusive i.e. the presence of Skewness implies absence of Kurtosis and vice-versa.

3.3 Tests of Skewness :

There are certain tests with the help of which it is possible to ascertain whether Skewness does or does not exist in a frequency distribution. They are :

%4 u In a skewed distribution values of Mean, Median and Mode would not coincide. The values of Mean and Mode are pulled away and the value of Median will be at the centre. In this distribution, Mean - Mode = 2/3 (Median - Mode).

%4 u Quartiles will not be equidistant from Median.

%4 u When the asymmetrical distribution is drawn on the graph paper, it will not give a bell shaped curve.

%4 u The sum of the positive deviation from the Median is not equal to sum of negative deviations.

%4 u Frequencies are not equal at points of equal deviations from the Mode

3.4. Nature of Skewness :

Skewness can be positive or negative or zero.

%4 u When the values of Mean, Median and Mode are equal, there is no Skenwness.

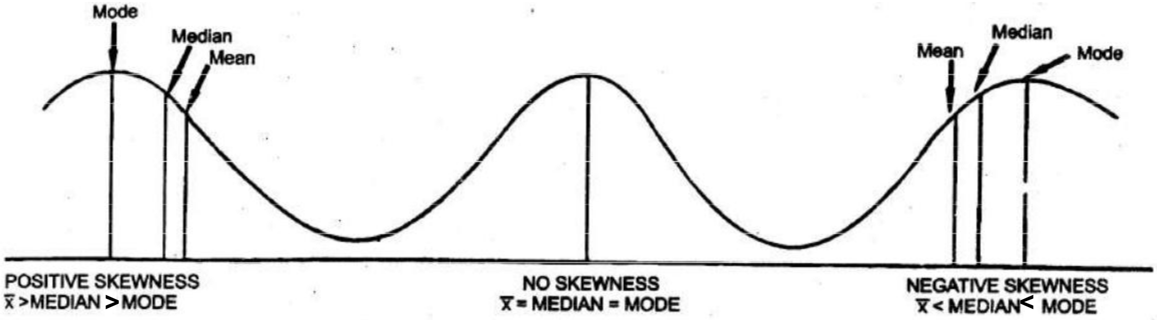
%4 u When the Mean > Median > Mode, Skewness will be positive.

%4 u When Mean < Median < Mode, Skewness will be negative.

Methods of finding out Skewness :

Skewness can be studied (a) graphically, (b) Mathematically. When we study Skewness graphically, in that case we observe, whether Skewness is positive or zero. This can be shown with the help of the following diagram:

The only limitation of this method is that we can simply study whether the Skewness is positive or negative or zero. We cannot find out any value of coefficient of Skewness.



In the mathematical methods, Skewness can be of two types:

%4 u Absolute Skewness.

%4 u Relative or coefficient of Skewness.

When the skewness is presented in absolute item i.e., in units, it is absolute Skewness. If the value of Skewness is obtained in ratios or percentages, it is called Relative or coefficient of Skewness.

When Skewness is measured in ‘absolute terms’ in that case we can compare one distribution with the other, if the units of measurements are different. When it is presented in ratios or percentages, comparison becomes easy. Relative measures of Skewness give coefficient of Skewness.

Characteristic of good measure of Skewness :

%4 u It should be a pure number in the sense that its value should be independent of the units of the series and also degree of variation in the series.

%4 u It should have a ‘zero-value’, when the distribution is symmetrical.

%4 u It should have a meaningful scale of measurement. So that we easily interpret the measured value. Mathematical measures of Skewness can be calculated on the basis of:

(a) Bowleys Method

(b) Karl-Persons’s Method

(c) Kelly’s Method

%4 u (a) Boweley’s Method :

Bowley’s method of Skeeness is based on the values of Median, lower and upper quartiles. This method suffer from the same limitations which are in the case of Median and quartiles.

Absolute Skewness

= $Q_3 + Q_1$ - 2 Median

Co-efficient of Relative Skewness

$$\frac{Q_3 \quad Q_1 \quad 2Median}{Q_1 Q_3}$$

The value of the coefficient in this method lies in between ± 1. This method is quite convenient for determining Skewness where one has already calculated quartiles.

For example: If the class intervals and frequencies are given as follows:

Class Intervals	Frequencies
	Frequency
Below 10	5
10—20	10
20—30	15
30—40	10
40 — above	5

In this case, if we want to calculate, 'coefficient of Skewness' on the basis of this methods, then we are required to calculate the values of Median Q_3 and Q_1 as follows:

Calculation of Co-efficient of Skewness on the basis of Median and Quartiles

Class Intervals	Frequencies	Cumulative Frequency
Below 10	5	5
10—20	10	15
20—30	15	30
30—40	10	40
40 — above	5	45

Calculations:

$$\text{Median} = l + \frac{n/2 \quad cf}{f} \quad i$$

$n/2 = \frac{45}{2}$ = 22.5, it lies in the cumulative frequency 30, corresponding to class (20 30)

$$\text{Median} \quad 20 \quad \frac{22.5 \quad 15}{15} \, 10 \quad 20 \quad \frac{7.5}{15} \, 10 \quad 25$$

$$Q_1 \quad l \quad \frac{n/4 \quad cf}{f} \, i$$

$n/4 \quad \frac{45}{4}$ = 11.25, lies in the cumulative frequency 15, corresponding to class interval (10 20).

Absolute Skewness: ($Q_3 + Q_1$ - 2 median).

Where Q_3 = 33.75 Q_1 = 16.25

Median = 25

Ab. Skewness = 33.75 +16.25 - 2(25) = 50 - 50 = 0

$$Q_1 \quad 10 \quad \frac{11.25-5}{} \quad 10 \quad 16.25$$

$$Q_3 \quad L \quad \frac{3n/4 - cf}{} \quad i$$

3n/4 = 33.75, that lies in the cumulative frequency 40, corresponding to group (30 40)

$$Q_3 \quad 30 \quad \frac{33.75-30}{} \quad 10 = 33.75$$

Co-efficient of Skewness

$$\frac{Q_3 \quad Q_1 \quad 2\text{Median}}{Q_3 \quad Q_1}$$

where Q_1 = Quartile or lower quartile,

Q_3 = Upper quartile,

In this problem,

Q_3 = 33.75,

Q_1 = 16.25, Median = 25

Co-efficient of Skewness $\frac{33.75 + 16.25\ 2(25)}{33.75-16.25} \quad 0 \ \frac{0}{17.50}$

Suitability:

Whether positional measures are called for, Skewness should be measured by Bowley's method. This method is, thus, helpful in the case of 'open-end series', where the importance of extreme value is ignored.

3.5. (b) Karl pearson's Method (Pearsonian Coefficient of Skewness)

Karl Pearsons has suggested two formulae:

%4 u where the relationship of Mean and Mode is established;

%4 u where the relationship between mean and Median is established.

The formulae that are devised by Karl-Pearson are when the value of Mean and Mode are related.

%4 u Absolute Skewness = Mean - Mode

%4 u Co-efficient of Skewness = X - Mode

%4 u

where = Standard deviation

¯%4 u = Simple Mean

THESE VALUES USUALLY LIE IN BETWEEN ± 1

When the value of Mean and Median are related

1. AB. Skewness = 3 (Mean - Median)

2. Co-efficient of Skewness $\frac{3(\text{Mean Median})}{} \quad \frac{\text{Mean Mode}}{}$

THE VALUES BYTHIS METHOD USUALLY LIES IN BETWEEN ± 3

Calculation of Coeffcient of Skewness on the Basis of Karl Pearson method by Taking:

Co-efficient of Skewness $\frac{\text{Mean Mode}}{}$

where s = Standard deviation

X =12, 18, 18, 22, 35, N=5.

$\bar{X}\ \ \frac{X}{N}\ \ \frac{105}{5}$ = 21 Marks.

Mode = 18 marks

X_2 Where x = (X - X)

N

N Solution :

Calculation of Co-efficient of Skewness

X	X-21 = x	X_2
12	— 9	81
18	— 3	9
18	— 3	9
22	+ 1	1
35	+ 14	196
N = 5		x_2 = 296

$\sqrt{\frac{296}{5}}\ \ \sqrt{59.2}\ \ 7.7$

Co-efficient of Skewness $\ \ $ **Mean Mode**

where mean = 21, Mode = 18, standard Deviation = 7.7

Co-efficient of Skewness $\frac{21\ \ 18}{7.7}\ \ \frac{3}{7.7}\ 4$

Calculation of Karl-Pearson's coefficient of Skewness by taking :

Coefficient of Skewness $\frac{3(\text{Mean}\ \ \text{Median})}{\ }$

In this given data = 12,18,18,12,35

Mean = 21, Median =18, = 7.7

Coefficient of skewness $\frac{3(21\ \ 18)}{7.7}\ \ \frac{3\ \ 3}{7.7}\ \ \frac{9}{7.7}$ 1.12

PROBLEMS

Exercise 1. Calculate the appropriate measure of Skewness from the following income-distribution:

Monthly Income (in Rs.)	Frequency
upto 100	9
101 — 150	51
151—200	120
201 — 300	240
301—500	136
501—750	33
751 — 1000	9
Above 1000	2
	N=600

Solution : In this problem, the open-ends series is given with the inclusive class-intervals. In this case Bowley's measure of Skewness is better, because it is based on Quartiles and Median. Moreover, Median and Quartiles do not take into account the extreme class-intervals.

Calculation of co-efficient of Skewness based on Quartile and Median.

Monthly Income (in Rs.)	Frequency	Cumulative Frequency
upto 100	9	9
101 — 150	51	60
151—200	120	180
201—300	240	420
301 — 500	136	556
501—750	33	589
751 — 1000	9	598
Above 1000	2	600
N=600		

Co-efficient of Skewness $\frac{Q_3 \quad Q_1 \quad 2\text{Median}}{Q_3 \; Q_1}$

Median = l $\frac{\frac{N}{2} \text{ c.f.}}{f}$ i

$\frac{N}{2} \; \frac{600}{2}$ 300; in the cumulative frequency 420, which is corresponding to group 201 — 300.

But the real limits of this class-interval are 200.5 — 300.5.

Median 200.5 $\frac{300 \; 180}{240}$ 100 200.5 Rs.250.5

Q_1 = l $\frac{\frac{N}{4} \text{ c.f.}}{f}$ i

n/4 = $\frac{600}{4}$ - 150. It lies the cumulative frequency 80, which is corresponding to class-interval 151 — 200. But the real limits of the class interval are 200.5 — 300.5

Q_1 150.5 $\frac{150 \;\; 60}{120}$ 50 150.5 37.5 Rs.188

Q_3 = L $\frac{\frac{3N}{4} \text{ c.f.}}{f}$ i

when 3n/4 is meant to find out upper quartile group.

3n/4 = $\frac{150 \;\; 60}{}$ 450

It lies in the cumulative 556, which is corresponding group 301 — 500. The real limits of this class interval are 300.5 — 500.5.

$$Q_2 = 300.5 \quad \frac{450\text{-}420}{136} \; 200 = 300.5 + \frac{30}{136} \; 200$$

$$= 300.5 \quad \frac{450\text{-}420}{136} = 300.5 + 44.12 \;\; \text{Rs.}344.62$$

Hence co-efficient of Skewness $\dfrac{Q_3 \quad Q_1 \quad 2\text{Median}}{Q_3 \quad Q_1}$

$$= \frac{344.62+188\text{-}2(250.5)}{344.62 \quad 188} \quad \frac{31.62}{156.62} \; 0.2 \text{ approx.}$$

Exercise 2. Calculate the appropriate measure of skew-ness from the following cumulative frequency distribution:

Age (under years)	No. of Persons
20	12
30	29
40	48
50	75
60	94
70	106

Solution: In this problem, we are given the upper limits of the class-intervals alongwith the cumulative frequency.

Therefore, we have to find out to lower limits and frequencies with the given data.

Formation of a frequency Distribution and Calculation of co-efficient of Skewness

Age (Years)	Number of Persons Frequency f	Cumulative Frequency c.f.
Below 20	12	12
20—30	17	29
30—40	19	48
40—50	27	75
50—60	19	94
60—70	12	106
	N=106	

In this case, the lower limits of first group is not given therefore, the best measure of Skewness is Bowley's Method. It is based on Quartiles and Median and does not take into account the extreme class-intervals.

Bowley's co-efficient Skewness $\dfrac{Q_3 \quad Q_1 \quad 2\text{Median}}{Q_3 \; Q_1}$

Thus, we have to calculate the values of Q_3, Q_1 and median.

$$\text{Median} = l \; \frac{\frac{N}{2} \; \text{c.f.}}{f} \quad i$$

Median has $\dfrac{\%4 \quad u}{2}$ items or $\dfrac{106}{2}$ or 53 items below it.

Therefore, it lies in the cumulative frequency 75, which is corresponding to the class-interval (40 — 50).
Hence, Median group is (40 — 50).
Here L = 40, i=10,

$f \quad 27, \frac{N}{2} \quad 53, c.f. \quad 48.$

Median = 40 + $\frac{53 \quad 48}{27}$ 10 =40+ $\frac{5}{27}$ 10 40 1.9 41.9

$Q_1 = l \frac{\frac{N}{4} \; c.f.}{f} \; i$

Q_1 has $\frac{\%4}{4}$ or $\frac{u \quad 106}{4}$ or 26.5 items below it.

It lies in the cumulative frequency 29, which is corresponding to the class interval 20 - 30.
Therefore, Q_1 group is 20 - 30.

Here l = 20 + $\frac{N}{4}$ 26.5, i = 10, c.f. = 12, f = 17

$Q_1 = 20+\frac{26.5\text{-}12}{17} \; 10$ $\quad Q_1 = 20+\frac{14.5}{17} \; 10 \quad 20 \quad 8.52 \quad 28.53$

$Q_3 = l \quad \frac{3N\text{-}c.f.}{f} \quad i$

Q_3 has $\frac{3N}{4}$ or $\frac{3 \quad 106}{4}$ or 79.5 items below it.

It lies in the cumulative frequency 94, which is corresponding to the group 50 - 60.
Therefore, Q_3 group is (50 - 60).

$Q_3 = 50+\frac{79.5\text{-}75}{19} \; 10$

or $Q_3 = 50+\frac{4.5}{19} \; 10 \quad 50 \quad 2.37 \quad 52.37$

Co-efficient of Skewness $\frac{Q_3 \quad Q_1 \quad 2Median}{Q_3 \quad Q_1}$

Here Q_3 = 52.37, Q_1 = 28.53, median = 41.9
Co-efficient of Skewness

$= \frac{52.37+28.53\text{-}2(41.9)}{52.37 \quad 28.53} \quad \frac{80.90\text{-}83.8}{23.84} \quad = \frac{-2.90}{23.84} 12.$

Therefore co-efficient of Skewness = - 12.

Excrcise 3. Calculate the Karl-Pearson's co-efficient of Skewness from the following data:

Mark (above)	No. of Students
0	150
10	140
20	100
30	80
40	80
50	70
60	30
70	14
80	0

Solution: In this problem, the lower limits of the class-intervals given corresponding to the cumulative frequencies.

Therefore, we have to find out the upper-limits and frequencies for the different groups.

Calculation of Co-efficient of Skewness

Marks	Number of Students	Cumulative
	(frequency)	Frequency
	f.	c.f.
0—10	10	10
10—20	40	50
20—30	20	70
30—40	0	70
40—50	10	80
50—60	40	120
60—70	16	136
70—80	14	150
	N=150	

In this problem, there are two model groups.

Therefore, the value of mode is ill-defined.

This formula of Karl-Pearson is applied to find out co-efficient os Skewness.

$$\frac{3(\text{Mean Median})}{\ }$$

Here $\text{Median} = l\ \frac{\frac{N}{2}\ \text{c.f.}}{f}\ \ i$

Median has $\frac{N}{2}$ or $\frac{150}{2}$ or 75 items below it.

It lies in the cumulative frequency 80, which is corresponding to the group (40 — 50).
Hence Median group is (40 to 50).

Here $L = 40$, $\frac{N}{2} = 75$, $f = 10$, c.f. $= 70$, i. 10.

Median $= 40\frac{75\text{-}70}{10}\ 10\ \ 45.$

Calculation of Mean and Standard-Deviation

Marks X	Frequency f	Mid points mean 45 X	Deviation from assumed	i = 10 d	d_2	fd	fd_2
0—10	10	5	-40	-4	16	-40	160
10—20	40	15	-30	-3	9	-120	360
20—30	20	25	-20	-2	4	-40	89
30—40	0	35	-10	-1	1	0	0
40—50	10	45	0	0	0	0	0
50—60	40	55	+ 10	+ 1	1	40	40
60—70	16	65	+ 20	+ 2	4	32	64
70—80	14	75	+ 30	+ 3	9	42	120
	N=150					fd = -86	fd_2 = 830

$\bar{\ }\%4 \quad u \ A \quad N\frac{fX}{}\ \ i$

Here A = 45, N = 150, i = 10. fd = -86

$\bar{X} \quad 45 \quad \frac{86}{150}\ 10$ or $45 - 5.73 = 39.27$

$$\sqrt{\frac{fd^{2}}{N} \quad \frac{fd}{N}^{2}}\ i$$

Here N = 150, i = 10, fd_2 = 830, fd = - 86

$$\sqrt{\frac{830}{150} \quad \frac{86}{150}^{2}}\ 10$$

or $\sqrt{5.5333\ \ (\ 57)^{2}\quad 10}$

or $\sqrt{5.5333\ \ .3249\ \ 10}$

$\sqrt{5.2048\ \ 10}$ or $2.28\ \ 10\ \ 22.8$

Co-efficient of Skewness $\frac{3(\text{Mean}\ \ \text{Median})}{}$

Here = 22.8, Mean = 39.27 Median = 45

Co-efficient of Skewness

$= \frac{3(39.27\text{-}45.00)}{22.8}$

or Co-efficient of Skewness

$$\frac{3(-5.73)}{22.8} = \frac{-17.19}{22.8} 0.754$$

Hence co-efficient of Skewness = - 0.754.

Exercise 4. (a) In a frequency distribution the coefficient of Skewness based on qualities is 0.6. If the sum of the upper and lower quartiles is 100 and Median is 38, find the values of lower and upper quartiles. Also find out the value of middle 50% items.

%4 u In a certain distribution, the following results were obtained:

Co-efficient of variation =

40% $\bar{X}$ **=25**

Mode = 20

Find out the co-efficient of Skewness by applying

$$\frac{\text{Mean} \quad \text{Mode}}{\text{Standard deviation}}$$

Solution : (a) Since Bowley's method is based on quartiles. Therefore, the formula of this basis is :

Co-efficinet of Skewness $\dfrac{Q_3 \quad Q_1 \quad 2\text{Median}}{Q_3 Q_1}$

Here co-efficient of Skewness = + 0.6

Median = 38, ($Q_3 + Q_1$) = 100

By substituting the values in the formula, we get $+ 0.6 = \dfrac{100\text{-}2(38)}{(Q_3 \quad Q_1)}$

By cross multiplying, we get : .6 ($Q_3 + Q_1$) = 100 - 76 = 24

$Q_3 \quad Q_1 = \dfrac{24}{6} = 40$

We are able to get, the equations and by adding, we get:

$Q_3+Q_1=100$ ....(i)

Q_3 --- Q_1 = 40 ....(ii)

$2Q_3 = 140$

$Q_3 \quad \dfrac{140}{2} \; 70$

Since $Q_3 + Q_1 = 100$

Q_1=100-70=30.

Hence the lower and upper quartiles are 30 an 70.

The value of middle 50% items can be obtained with the help of ($Q_3 + Q_1$)

The value of middle 50% item is (70-30) = 40

%4 u In this problem, the value of standard-deviation is missing. This we can calculate by applying the following formula:

C.V. = $\bar{X}$ **100**

Here C.V. = 40%, $\bar{X}$ **= 25 are given.**

$40 = \overline{25}\ 100$

or $40 = 4$ or $= \frac{40}{4} = 10$

Co-efficient of Skewness $= \frac{(\text{Mean Mode})}{}$

Here Mean = 25, Mode = 20, = 10

Coeff. of Skewness $= \frac{25\ 20}{20} = .5$

Hence co-efficient of Skewness = + .5

Exercise 5. What is the relationship between Mean, Median and Mode in:

%4 u Symmetrical curve.

%4 u A negatively skewed curve.

%4 u A positively skewed curve.

From the following marks obtained by 120 students each in section A and B of a class, the following measures are secured :

Section A	Section B
Mean = 47 marks	Mean = 48 marks
Standard deviation	Standard deviation
= 15 marks	=15 marks
Mode =52 marks	Mode = 45 marks.

Find out the co-efficient of Skewness and deviative the degree of Skewness and in which distribution, the marks are more skewed.

Solution : The relationship between Mean, Median and Mode in different cases, can be established as :

In a symmetrical curve, there is no question of Skewness. Therefore, in this case, the value of Mean = Median

%4 uMode.

In a negatively skewed curve, the value of mean is less than median is less than mode. In other words, Mean < Median < Mode.

In a positively skewed curve, the value of Mean is greater than Mode. In other words, Mean > Median > Mode. In the given problem, for finding out the degree of Skewness, we have to compute the co-efficient of Skewness. Calculation of Co-efficient of Skewness

In Sections “A” and “B”

[Section A]

Co-efficient of Skewness $\frac{\text{Mean Mode}}{}$

Here $\bar{X}$ = 47, Mode = 5.2

=15

Therefore, co-efficient of Skewness

$\frac{47\ \ 52}{15} \ \frac{5}{15} \ 0.33$

Hence the marks are negatively skewed

[Section B]

Co-efficient of Skewness = $\frac{\text{Mean Mode}}{\quad}$

Here $\bar{X}$ = 48, Mode = 45

=15

Therefore, co-efficient of Skewness

$\frac{4845\ 3}{15\,15}$ .2 —

Hence the distribution of marks is positively skewed.

On comparison, the distribution of marks in 'Section A' is more skewed as compared to 'Section B'.

%4 u Self-

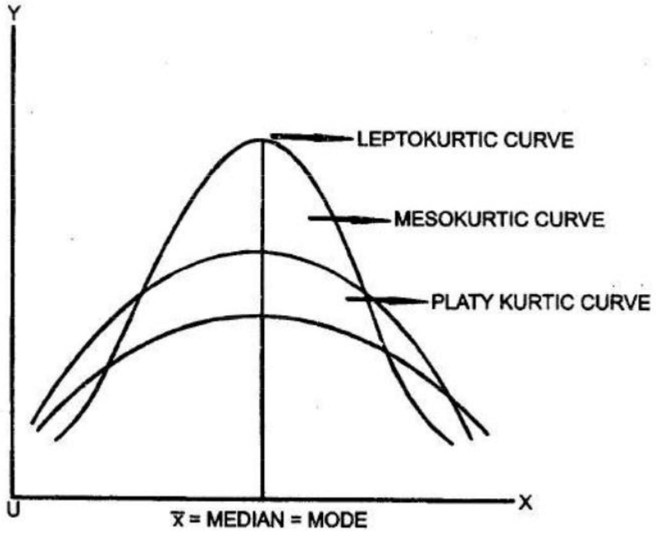
Assessment Fill in

Blanks:

1. If Q3 = 30, Q1 =20, Med =25, Coeff. of Sk. shall be ....
2. If Coeff. of Sk. = 0.8, Median = 35, = 12, the mean shall be ....
3. In a symmetrical distribution the coefficient of skewness is ....
4. The limits for Bowley's coefficient of skewness are ....

3.6. KURTOSIS

This is another characteristics used for description and comparisons. It should the extent to which the curve is more peaked or more flat-topped than the national curve. The stand point of Kurtosis, the normal curve is mesokurtic, i.e., of "intermediate peakedness".When the curve of a distribution is relatively flatter than the normal curve, it is said to have Kurtosis. When the curve or polygon is relatively more peaked, 'it is said to lack Kurtosis. Thus the concept of Kurtosis helps us in studying the peakedness of the distribution.



where $_2$ = 3, Mesokurtic Curve

$_2$ < 3, Platykurtic Curve

$_2$ > 3, Leptokurtic Curve

Measures of Kurtosis :

Measures of kurtosis is denoted by $_2$ and in a normal distribution $_2$ =3.

If $_2$ is greater than 3, the curve is more peaked and is known as leptokurtic, if $_2$ is less than 3, the curve is flatter at the top than the normal, and is known as plary-kurtic. Thus Kurtosis is measured by

$$\beta_2 \quad \frac{\mu^1}{\mu_2{}_2} \qquad \frac{\dfrac{fx^4}{n}}{\dfrac{fx^{4\;2}}{n}}$$

where x = (X - $\bar{X}$)

R.A. Fisher has introduced another notation Greek letter gamma, symbolically,

$$y^2 \;\; \beta^2 \;\; 3\,\beta \qquad 2 \;\; \frac{\mu^4}{\mu^2{}_2} \; 3.$$

In this case of a normal distribution, y_2 is zero. If y is more than zero. (i.e. positive) then the curve is platykurtic

and if y_2 is less than 0 (i.e. negative) then the curve is leptokurtic. It may be noted that $\mu_4 \;\; \frac{fX^4}{N}$ is an absolute measure of Kurtosis, but $\beta_2 \frac{4}{2_2}$ is a relative measure of Kurtosis. Also bigger the value of y_2 in a frequency distribution, the greater is its departure from normality.

Skewness and Kurtosis, $_1$ and $_2$ are statistical parameters which determine whether a given distribution is normal or not. $_1$ and $_2$ are measures of symmetry and normality respectively. If $_1$ = 0, the distribution is symmetrical and if, in addition $_2$ = 3, the distribution is normal or not. $_1$ and $_2$ are measures of symmetry and normality respectively. If $_1$ = 0, the distribution is symmetrical and if, in addition $_2$ = 3, the distribution is normal.

3.7. Comparison between dispersion, Skewness and Kurtosis :

Dispersion, Skewness and Kurtosis are the different characterstics of frequency distribution. Dispersion studies the scatter of the items round a central value or among themselves. It does not show the extent to which deviations cluster below an average or above it. This is studies by Skewness. In other words, this tells us about the cluster of the deviations above and below a measure of central tendency. Kurtosis studies the concentration of the items at the central part of a series. If the items concentrate too much in the centre, the curve becomes 'LEPTOKURTIC' and if the concentration in the centre is comparatively little, the curve becomes 'PLATYKURTIC'.

Exercise 1. From the following data find out the value of B_2 and also study the nature of Kurtosis..

X	f
0—10	6
10—20	10
20—30	15
30—40	10
40—50	5
	Total = 45

Solution :

Calculation of B_2

X Classes	f Frequency	Mid Points X	fx	X	X^2	X^3	X^4	fx_2	fx_3	fx_4
0-40	5	5	25	20	400	8000	160000	2000	40000	800000
10—20	10	15	150	10	100	1000	10000	1000	10000	100000
20—30	15	25	375	0	0	0	1	6	0	0
30—40	10	35	350	+ 10	100	1000	10000	1000	1000	100000
40—50	5	45	225	+ 20	400	8000	160000	4000	40000	800000
	N=45	fx =					fx_2 =		fx_3=0	fx_4 =
		1125						6000		1800000

Calculation of Mean $\frac{fX}{N}$ $\frac{1125}{45}$ = 25

Calculation of $B_2 \; \frac{4}{2^2}$

where $\mu_4 \; \frac{fX^4}{N} \, 40000$

$$\mu_2 \; \frac{fX^2}{N} \quad \frac{6000}{45} \quad 40000$$

$$\beta_2 \; \frac{4}{2^2} \quad \frac{40000}{(133.33)^2} \quad 3$$

Since the value of B_2=3

therefore the distribution is MESOKURTIC.

(B) Self Exercise

%4 u What do you mean by skewness?

%4 u What are the methods of finding out skewness?

%4 u What is the difference between dispersion, skewness and kurtosis?

%4 u Summary:

Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution. Measures of skewness tall us the direction and extent of asymmetry in a series, and allow us to compare two or more series. Measures of skewness may be absolute or relative.

Kurtosis refers to the flatness or peakedness of the curve. Measure of Kurtosis tell us the extent to which a distribution is more peaked or flat-topped than the normal curve.

3.9 GLOSSARY:

Skewness: It is a measure that studies the degree of departure from symmetry.

Kurtosis- It referes to the flatness or peakedness of the curve. It is a meausre of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

3.10 Answers: Self Assessment

1. 0 2. 108.2 3. Zero 4 ± 1

%4 u Refer to section 3.2

%4 u Refer to section 3.5

%4 u Refer to secton 3.7

3.11. Terminal Questions

%4 u What is skewness? Describe the Various measures of skewness.

%4 u Distinguish between dispersion and skewness.

%4 u Find the coefficient of skewness from the following information:

Difference of two Quartiles	**= 8**
Mode	**= 11**
Sun of two Quartiles	**= 22**
Mean	**= 8**

%4 u What is Kurtosis? How is it measured?

%4 u 12. Suggested Readings

%4 u Heinz Kohler, Statistics for Business and Economics, Harper Collins.

%4 u Hooda, R.P., Statistics for Business and Economics, MacMillan, New Delhi.

%4 u Gupta, S.P., Statistical Methods, Sultan Chand & Sons, New Delhi.

*****

Lesson-4
Correlation Analysis

Structure:

4.1 Learning Objectives:

After studying the lesson, you should be able to understand:

%4 u **Meaning and Utility of Correlation.**

%4 u **Various Types of Correlation.**

%4 u **What are the different methods of studying Correlation.**

%4 u **Introduction:**

In the earlier chapters we have discussed univeriate distributions to highlight the important characteristics by different statistical techniques. Univariate distribution means the study related to one variable only. We may however come across certain series where each item of the series may assume the values of two or more variables. The distribution in which each unit of series assumes two values is called bivariate distribution. In a bivariate distributions, we are interested to find out whether there is any relationship between the two variables. The correlation is a statistical technique which studies the relationship between two or more variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of relationship between the two variables. When two 'variables are related in such a way that a change in the value of one is accompanied either by a direct change or by an inverse change in the values of the other, the two variables are said to be correlated. In the correlated variables an increase in one variable is accompanied by an increase or decrease in the other because keeping other things equal, relationship between the price and demand of a commodity shall cause a decrease in the demand for that commodity. Relationship might exist between the heights and weights of the students and between amount of rainfall in a city and the sales of raincoats in that city.

4.3. Definition:

These are some of the important definitions about Correlation.

Croxton and Cowden says, "When the relationship is of a quantitative nature, the appropriate statistical tool; for discovering and measuring the relationship and expressing it in a brief formula is known as Correlation".

A.M. Tuttle says, "Correlation is an analysis of the covariation between two or more variables."

W.A. Neiswanger says, "Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective.

L.R. Connor says, "If two or more quantities vary in sympathy so that the movements in one tends to be accompanied by corresponding movements in others than they are said to be Correlated.

4.4. Utility of Correlation :

The study of correlation is very useful in practical life as revealed by these points.

%4 u **With the help of Correlation analysis, we can measure in one figure, the degree of relationship existing between variables like price, demand, supply, income, expenditure etc. Once we know that two variables are correlated then we can easily estimate the value of one variables, given the value of other.**

%4 u Correlation analysis is of great use of economists and businessmen, it reveals to the economists the disturbing factors and suggest to him the stabilizing forces. In business, it enables the executive to estimate costs, sales etc. and plan accordingly.

%4 u Correlation analysis is helpful to scientists. Nature has been found to be a multiplicity of inter-related forces.

%4 u Difference between Correlation and Causation :

The term correlation should not be misunderstood as causation. If correlation exists between two variables, it must not be assumed that a change in one variable is the cause of a change in other variable. In simple words, a change in one variable may be associated with a change in another variable but his change need not necessarily be the cause of a change in the other variable. When there is no cause and effect relationship between two variables but a correlation is found between two variables such correlation is known as "superious correlation" or "nonsense correlation". Correlation may exist due to the following:

%4 u Pure change Correlation : This happens in a small sample. Correlation may exist between incomes and weights of four persons although there may be no cause and effect relationship between income and weights of people. This type of correlation may arise due to pure random sampling variation or because of the bias of investigator in selecting the sample,

%4 u When the correlation variables are influenced by one or more variables. A high degree of correlation between the variables may exist, where the same cause is affecting each variable or different cause affecting each with the same effect. For instance, a degree of correlation may be found between yield per acre of rice and tea due to the fact that both are related to the amount of rainfall but none of the two variables is the cause of the other.

%4 u When the variable mutually influence each other so that neither can be called the cause of other. At times it may be difficult to say that which of the two variables is the cause and which is the effect because both may be reacting on each other.

%4 u Types of Correlation :

Correlation is described in these four ways ;

%4 u Positive and Negative

%4 u Simple and Multiple.

%4 u Partial and Total

%4 u Linear and Non-Linear

(Curvilinear) Positive and Negative

Correlation :

Positive or direct Correlation refers to the movement of variables in the same direction. The correlation is said to be positive when the increase (decrease) in the value of one variable is accompanied by an increase (decrease) in the value of other variable also. Negative or inverse correlation refers to the movement of the variables in the opposite direction. Correlation is said to be negative, if an increase (decrease) in the value of one variable is accompanied by a decrease (increase) in the value of other.

Sample and Multiple Correlation :

Under simple correlation, we study the relationship between two variables only i.e., between the yield of wheat and the amount of rainfall or between demand and supply of a commodity. In case of multiple correlation, the relationship is studied among three or more variables. For example, the relationship of yield of what may be studied with both chemical fertilizers and the pesticides.

Partial and Total Correlation :

There are simply two categories of multiple correlation analysis. Under partial correlation the relationship of two or more variable is studied in such a way that only one dependent variable and one independent variable is considered and all others are kept constant. For example, coefficient of correlation between yield of wheat and chemical fertilizers excluding the effects of pesticides and manures is called partial correlation. The total correlation is based upon all the variables.

Linear and Non-Linear Correlation :

When the amount of change in one variable tends to kept a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. But if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable then correlation is said to be non-linear. The distinction between linear and non-linear is based upon the consistency of the ratio of change between the variables.

4.7. Methods of Studying Correlation :

These are the methods which help us to find out whether the variables are related or not:

%4 u Scatter Diagram Method.

%4 u Graphic Method.

%4 u Karl Pearson's Coefficient of Correlation.

%4 u Rank Method.

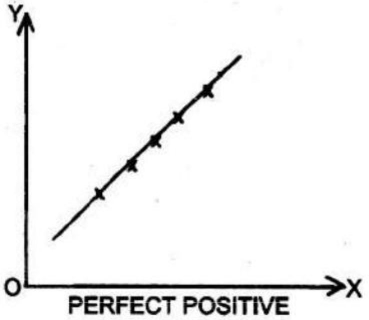
%4 u Concurrent deviation method.

We shall discuss these method in detail now.

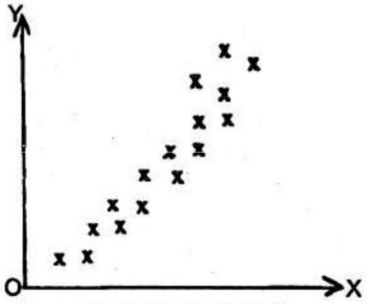
%4 u Scatter Diagram : Scatter diagram is drawn to visualize the relationship between two variables. The values of more important variable is plotted on the X-axis while the values of the other variable are plotted on the Y-axis. On the graph, dots are plotted to represent the various pairs of figure. When dots have been plotted to represent all the pairs, we get a scatter diagram. The way the dots scatter gives an indication of the kind of relationship which exists between the two variables.

When there is a positive correlation between the variables, the dots on the scatter diagram run from left hand bottom to the right hand upper corner. In case of perfect positive correlation all the dots will lie on a straight line.

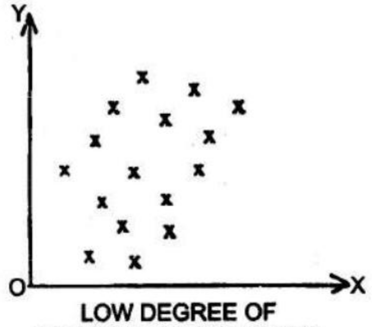
Note : While drawing scatter diagram, it is not necessary to take at the point of sign the zero values of X and Y variables, but the minimum values of the variables under consideration may be taken.



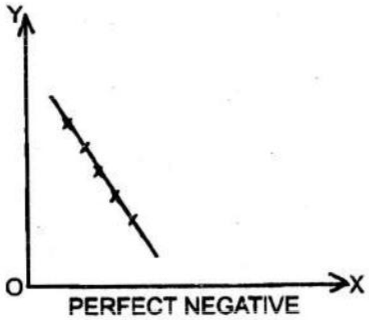
PERFECT POSITIVE CORRELATION



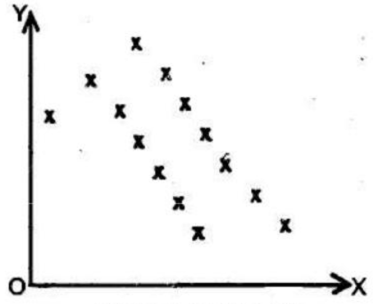
HIGH DEGREE OF POSITIVE CORRELATION



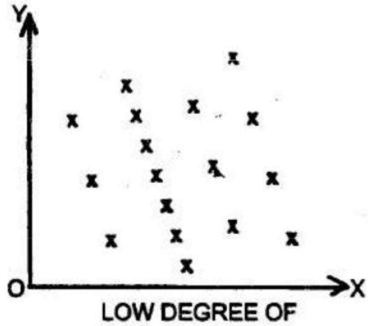
LOW DEGREE OF POSITIVE CORRELATION



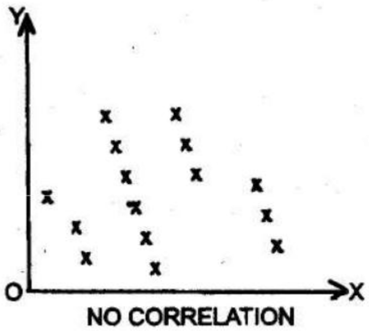
PERFECT NEGATIVE CORRELATION



HIGH DEGREE OF NEGATIVE CORRELATION



LOW DEGREE OF NEGATIVE CORRELATION



NO CORRELATION

When a negative correlation exists between the variables, dots on the scatter diagram run from the upper left hand corner to the bottom right hand corner. In case of perfect negative correlation, all the dots lie on a straight line.

If a scatter diagram is drawn and no path is formed, there is no correlation. Therefore there is no correlation when all the dots are scattered over the graph without any system.

Students are advised to prepare two scatter diagrams on the basis of the following data :

(i) Data for the first Scatter Diagram :

Demand Schedule

Price (in Rs.)	Commodity Demand (in units)
6	180
7	150
8	130
9	120
10	125

(ii) Data for the second Scatter Diagram :

Supply Scheduly

Price (in Rs.)	Commodity Supply
50	2,000
51	2,100
52	2,200
53	2,500
54	3,000
55	3,800
56	4,700

Students will find that the first diagram indicate a negative correlation where the second diagram shall reveal a positive correlation.

%4 u Graphic Method. In this method the individual values of the two variables are plotted on the graph paper. Therefore two curve obtained—one for the X variables and another for the Y variable.

Interpreting Graph :

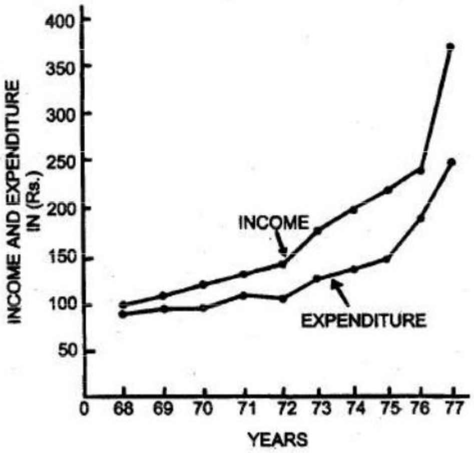
The graph is interpreted as follows :

%4 u If the curves run parallel or nearly or more in the same direction, there is positive correlation.

%4 u On the other hand, if the curves move in the opposite direction, there is a negative correlation.

Illustration 1. Show the correlation of the following data by the graphic method :

Year	Average Income (Rs.)	Average Expenditure (Rs.)
1968	100	90
1969	110	95
1970	125	100
1971	140	120
1972	150	120
1973	180	140
1974	200	150
1975	220	170
1976	250	200
1977	360	260



The graph prepared shows that income and expenditure have a close positive correlation. As the average income increases, the average expenditure also increases.

%4 u Karl Pearson's Coefficient of Correlation: Karl Pearson's method, popularly known as Pearsonian coefficient of correlation, is most widely applied practice to measure correlation. The Pearsonian coefficient of correlation is denoted by the symbol r.

According to Karl Pearson's method, coefficient of correlation of the variable is obtained by dividing the sum of the products of the corresponding deviations of the various items of two series from their respective means by the products of their standard deviations and the number of points of observations.

Symbolically,

$$r \quad \frac{xy}{N \ _{x} \ _{y}} \qquad .....(i)$$

where r stands for coefficient of correlation.

where x_1, x_2, x_3, x_4 ... ... x_n are the deviations of various items of the first variable from their mean y_1 , y_2, y_3, ... y_n are the corresponding deviations of the items of the second variable from their mean, xy is the sum of the products of these corresponding deviations. N stands for the number of pairs of items, $_x$ stands for the standard deviation of X variable and $_y$ stands for the standard deviation of Y variable.

$$_x \ \sqrt{\frac{x^2}{N}} \ \text{and} \quad _y \ \sqrt{\frac{y^2}{N}}$$

If we substitute the value of $_x$ and $_y$ in the above mentioned formula of computing r, we get

$$r \quad \frac{xy}{N\frac{\sqrt{x^2}}{\sqrt{N}} \quad \frac{\sqrt{y^2}}{\sqrt{N}}}$$

$$\text{or } r = \frac{xy}{\sqrt{x^2 \quad y^2}} \qquad \text{....... (ii)}$$

The degree of correlation varies between + 1 and -1 ; the result will be + 1 in case of perfect positive correlation and - 1 in the case of perfect negative correlation.

Computation of correlation coefficient can be simplified by dividing the given data by a common factor. In such a case, the ultimate result is not multiplied by the common factor because coefficient of correlation is independent of change of scale and origin.

Exercise 1

Calculate Co-efficient of Correction from the following data:

X	50	100	150	200	250	300	350
Y	10	20	30	40	50	60	70

Solution :

X	$X \quad \bar{X}$ (200)	$\frac{X \quad \bar{X}}{50}$ x	x_2 (40)	Y y	$Y \quad \bar{Y}$	$\frac{Y \quad \bar{Y}}{10}$	y_2	xy
50	150	3	9	10	30	3	9	9
100	100	2	4	20	20	2	4	4
150	50	1	1	30	10	1	1	1
200	0	0	0	40	0	0	0	0
250	+50	+1	1	50	+10	+1	1	1
300	+100	+2	4	60	+20	+2	4	4
350	+150	+3	9	70	+30	+3	9	9
		x = 0	x_2 = 28			y = 0	y_2 = 28	xy = 28

Now according to formula $r = \frac{xy}{\sqrt{x^2 \quad y^2}}$

By substituting the values we get $r = \frac{28}{\sqrt{28 \quad 28}}$ $r = \frac{28}{28}$ 1.

Thus we find that there is perfect positive Correlation.

Hence there is perfect positive Correlation.

Exercise 2

A sample of five items is taken from the production of a firm, length and weight of the five items are given below:

Length (inches)	3	4	6	7	10
Weight (ounces)	9	11	14	15	16

Calculate Karl Pearson's correlation coefficient between length and weight in the above sample and interpret the value of this coefficient

Solution : $\overline{X} = \frac{X}{N} \quad \frac{30}{N} \; 6$

$$\overline{Y} = \frac{Y}{N} \quad \frac{65}{5} \; 13$$

X	$(X \quad \overline{X})$ x	x_2	Y	$Y \quad \overline{Y}$	y_2	xy
3	3	9	9	4	16	12
4	2	4	11	2	4	4
6	0	0	14	+1	1	0
7	+1	1	15	+2	4	2
10	+4	16	16	+3	6	12
x = 30	0	30	y=65	0	34	30

$$r = \frac{xy}{\sqrt{x^2} \; y^2}$$

Where $xy = 30$
$x^2 = 30$
$y^2 = 34$

$$r = \frac{30}{\sqrt{30 \;\; 34}} \quad \frac{30}{\sqrt{1020}} \; 0.939 \text{ Ans.}$$

The value of r indicate that there exists a high degree of positive correlation between length and weights. Exercise 3

From the following data compute the coefficient of correlation between X and Y

	X-Series	Y-Series
Number of items	15	15
Arithmetic Mean	25	18
Square of deviation from Mean	136	138

Summation of product deviations of X and Y from their Arithmetic Means = 122.

Solution: Denotation deviations of X and Y from their Arithmetic Mean by x and y respectively, the given data is:

$x^2 = 136 \qquad xy = 122$
$x^2 = 138$

$$r = \frac{xy}{\sqrt{x^2} \;\; y^2}$$

$$r = \frac{122}{\sqrt{136 \; 138}} = \frac{122}{137} \; 0.89 \text{ Ans.}$$

Short-cut Method : To avoid difficult calculations due to mean being in fraction, deviations are taken from assumed means while calculating coefficient of correlation. The formula is also modified for standard deviations because deviations are taken from assumed means. Karl Pearson's formula for short-cut method is represented symbolically as follows:

$$r = \frac{dxdy - \frac{dx\ .\ dy}{N}}{\sqrt{dx^2 \quad \frac{(dx^2)}{N}}\, dy^2 \sqrt{\quad \frac{(dy^2)}{N}}}$$

or

$$r = \frac{N\ dx.dy - \ dx \quad dy}{\sqrt{N\ dx^2 \quad (\ dx)^2}\ \sqrt{N\ dy^2 \quad (\ dy)^2}}$$

Illustration 5

Compute the coefficient of correlation from the following data :

where :

Marks in Statistics	Marks in Mathematics
20	18
30	35
28	20
17	18
19	25
23	28
35	33
13	18
16	20
38	40

Solution :

Marks in Statistics X	X-30 dx	dx$_2$	Marks in Maths Y	Y-30 dy	dy$_2$	dxdy
20	10	100	18	12	144	+ 120
30	0	0	35	+ 5	25	0
28	2	4	20	10	100	+ 20
17	13	169	18	12	144	+ 156
19	11	121	25	5	25	+ 55
23	7	49	28	2	4	+ 14
35	+ 5	25	33	+ 3	9	+ 15
13	17	289	18	12	144	+ 204
16	14	196	20	10	100	+ 140
38	+ 8	64	40	+ 10	100	+ 80
N=10	61	1017		45	795	804

$$r = \frac{dxdy - \frac{dx.\ \ dy}{N}}{\sqrt{dx^2 \qquad \frac{(\ dx^{2})}{N}}\ dy^2\sqrt{\qquad \frac{(\ dy^{2})}{N}}}$$

where:-

dx deviation of X series from an assumed mean 30.

dy deviation of Y series from an assumed mean 30.

dx_2 sum of the squares of the deviations of X series from assumed mean.

dy_2 sum of the squares of the deviations of Y series from assumed mean.

dxdy sum of the product of deviations of X and Y series from their assumed means.

$$\text{Hence } r = \frac{804 - \frac{(\ 61)\ (\ 45)}{10}}{\sqrt{107 \quad \frac{(\ 61)^2}{10} 795 \quad \frac{(\ 45)^2}{10}}}$$

$$\text{or} \quad r = \frac{8040 \quad 2745}{\sqrt{(10170\ 3721)(7950 \quad 2025)}}$$

$$\text{or } r = \frac{5295}{\sqrt{6449\,5925}} \qquad r = \frac{5295}{\sqrt{3,82,10,325}} \qquad r = \frac{5295}{\sqrt{6181.45}} \qquad r = 0.86 \text{ Ans.}$$

Coefficient of Correlation for Continuous Series

In the case of continuous series, we assume that every item which falls within a given class interval falls exactly at the mid-value of that class. The formula, because of the presence of frequencies is modified as follows :

$$r = \frac{fdxdy - \frac{(\ fdx).(\ fdy)}{f}}{\sqrt{fdx^2 \qquad \frac{(\ fdx^{2})}{f}}\sqrt{fdy^2 \qquad \frac{(\ fdy)^{2}}{f}}}$$

The various values shall be calculated as follows :

%4 u Take the step deviation of variable X and denote by ax.

%4 u Take the step deviation of variable Y and denote by dy.

%4 u Multiply dx dy and the respective frequency of each cell and write the figure obtained in the right-hand upper corner of each cell.

%4 u Add all the concerned values calculated in step (iii) and we get fdxdy.

%4 u Multiply the frequencies of the variable X by the deviations of X and we get fdx.

%4 u Take the squares of the deviations of the variable X and multiply them by the respective frequencies and we get fdx_2.

%4 u Multiply the frequencies of the variable Y by the deviations of Y and we get fdy.

%4 u Take the squares of the deviations of the variable Y and multiply them by the respective frequencies and we get fdy_2.

%4 u Now substitute the value of fdxdy, fdx, fdx_2, fdy, fdy_2 in the formula and we get the value of r.

Illustration 6

The following table gives the ages of husbands and wives at the time of their marriages. Calculate the correlation coefficient between the ages of husbands and wives

Age of Husbands

	X Series / Y Series	20 30	30 40	40 50	50 60	60 70	Total
	15 25	5	9	3			
	25 35		10	25	2		37
Age s of wiv es	35 45		1	12	2		15
	45 55			4	16	5	25
	55 65				4	2	6
	Total	5	20	44	24	7	100

	Age groups				20-30	30-40	40-50	50-60	60-70				
	Mid Points				25	35	45	55	65				
	Deviation from Assumed mean 45				-20	-10	0	+10	+20	Total			
	Step deviation Common factor = 10				-2	-1	0	+1	+2	(f)	fdy	fdy^2	dxdy
	15 25	20	20	2	5	9	3			17	34	68	38
	25 35	30	10	1		10	25	2		37	37	37	8
Age of Wives (Y)	35 45	40	0	0		1	12	2		15	0	0	0
	45 55	50	+10				4	16	5	25	25	25	26
	155 65	60	+20	2				4	2	6	12	24	16
5	Ag e gr ou ps	Mid Poin ts	DeviationfromAssumedmean45Stepdeviation	Co m mo nfa cto r = 10	5	20	44	24	7	Ef=10	Efdy = 34	Efdy$_2$ 154	Efdxdy =88
					-10	-20	0	24	14	Efdx = 8			
					20	20	0	24	28	Efdx^2 =92			
					20	28	0	22	18	Efdxdy =88			

$$r = \frac{fdxdy - \frac{(\ fdx).(\ fdy)}{f}}{\sqrt{fdx_2 \quad \frac{(\ fdx^2)}{f} \quad fdy_2 \quad \frac{(\ fdy)^2}{f}}} = \frac{88 - \frac{(8)(34)}{100}}{\sqrt{92 \quad \frac{(8)^2}{100} 154 \quad \frac{(34)^2}{100}}}$$

$$= \frac{90.72}{\sqrt{91.36 \quad 142.44}} = 79.$$

Merits of Pearson 's coefficient of Correlation : The correlation coefficient summarizes in one figure not only the degree of correlation but also the direction. Value varies between + 1 and - 1.

Demerits of Pearson's coefficient of Correlation : It always assume linear relationship between the variables infact the assumption may be wrong. Secondly, it is not easy to interpret the significance of correlation coefficient. Also, the method is time consuming and is affected by the extreme items.

Probable Error of the coefficient of Correlation : It is calculated to find out how far the Pearson's coefficient of correlation is reliable a particular case, The formula is P.E. of coefficient of correlation

$$\%4 \quad u \quad \frac{.6745}{\sqrt{N}} \quad 1 \;\; r_2$$

where r = coefficient of correlation and

N = number of pairs of items.

If the probable error calculated is added to and subtracted from the coefficient of correlation, it would give us such within which we can expect the value of the coefficient of correlation to vary.

If r less than probable error, then there is no real evidence of correlation is considered highly significant.

If r is more than 6 times the probable error; the coefficient of real correlation.

If r is more than 3 times the probable error but less than 6 times, correlation is considered significant but not highly significant.

If the probable error is not much and the given r is more than the probable error but less than 3 times of it, nothing definite can be concluded.

%4 u Self Assessment Fill in the blanks :

1. **Distributions relating to a single characteristics are known ....**
2. **Study of 'Correlation' is meant to determine whether there exists some sort of .... between the variables.**
3. **.. the degree of association between two or more variables.**
4. **Correlation is an analysis of ....between two or more variables.**
5. **................................... is a numerical measure of the degree of association between two or more variables.**
6. **The coefficient of correlation lies between .... and ....**
7. **A .... of the data helps in having a visual idea about the nature of association between two variables.**

%4 u Rank Correlation : There are many occasions in problems of business and industry when it is not possible to measure the variable under consideration quantitatively or the statistical series is composed of items which can not be exactly measured. For instance, it may be possible for the two judges to rank six different brands of cigarettes in terms of taste, whereas it may be difficult to give them a numerical grade in terms of taste. In these type of problems, Spearman's coefficient of rank correlation is used to find out the relationship. The formula for rank correlation is

$$= 1- \frac{6 \;\; D^2}{N(N^2 \;\; 1)} \qquad \text{or} \qquad = 1- \frac{6 \;\; D^2}{N^3 \;\; 1}$$

where = stands for rank coefficient of correlation.

D = refers to the difference of ranks between paired items.

N = refers to the number of paired observations. The value of rank correlation coefficient also varies between + 1 and - 1. When the value of = + 1, there is complete agreement in the order of ranks and the ranks will be in the same order. When = - 1, the ranks will be in opposite direction showing complete disagreement in the order of ranks. Let us take an illustration.

Illustration 7

The ranking of 10 individuals at the start and as the finish of a course of training are as follows:

Individual	Rank before	Rank after
A	1	6
B	6	8
C	3	3
D	9	7
E	5	2
F	2	1
G	7	5
H	10	9
I	8	4
J	4	10

Calculate coefficient of correlation

Solution :

	R_1	R_2	D (R_1 - R_2)	D_2
A	1	6	-5	25
B	6	8	-2	4
C	3	3	0	0
D	9	7	-2	4
E	5	2	3	9
F	2	1	1	1
G	7	5	2	4
H	10	9	1	1
I	8	4	4	16
J	4	10	-6	36
N= 10			D_2 = 100	

By applying the formula

$$= 1- \frac{6\ D^2}{N^3\quad N} \qquad = 1- \frac{6\ \ 100}{10^3\quad\ 10}$$

$$= 1- \frac{600}{1000\ \ 10}\quad 1 \quad \frac{600}{990} \quad 1\ .609\ \ 0.394 \text{ Ans.}$$

When we are given the actual data and not the ranks, it becomes necessary for us to assign the ranks. Ranks can be assigned by taking either the highest value as one or the lowest value as one. But whether we start by taking the highest value or the lowest value we must follows the same order in case of both the variables.

Illustration 8

Calculate Rank Correlation from the following data:

X	:	17	13	15	16	6	11	14	9	7	12
Y	:	36	46	35	24	12	18	27	22	2	8

Solution :

Calculation of Rank Correlation

X	R_1	Y	R_2	D	D_2
	(Ranks)	(Ranks)	$(R_1 - R_2)$		
17	1	36	2	I	1
13	5	46	1	+ 4	16
15	3	35	3	0	0
16	2	24	5	3	9
6	10	12	8	+ 2	4
11	7	18	7	0	0
14	4	27	4	0	0
9	8	22	6	+ 2	4
7	9	2	10	1	1
12	6	8	9	3	9
N= 10					D_2 = 100

The Rank Correlation coefficient is calculated by

$$1 - \frac{6 \ D^2}{N_3 \quad N}$$

$$r = 1 - \frac{6 \ \ 44}{10^3 \quad 10} = 1 \ \ \frac{2640}{990}$$

$$\rho \quad 1 \ \ .266 \text{ or } \rho \quad 0.734 \text{ Ans.}$$

In some cases it becomes necessary to rank two or more than two individuals as equal. In these cases it is customary to give each individual an average rank. Therefore, if two items are equal for 4th and 5th rank, each item shall be

ranked as 4.5 i.e., $\frac{4 + 5}{2}$. It means, where two or more items are to be ranks equal, the rank assigned for purposes of calculating coefficient of correlation is the average of ranks which these items would have got had they differed slightly from each other. When equal ranks are assigned to some items, the rank correlation formula is also adjusted.

The adjustment consists of adding $\frac{1}{12}$ $(m_2 - m)$ to the value of D_2 where m stands for number of items whose ranks are common.

The formula can be written as

$$\rho = 1 \ \frac{6 \ D^2 + \frac{1}{12}(m^3 \quad m) \quad \frac{1}{12}(m^3 \quad m)\ldots \ \ldots}{N^3 \quad N}$$

Let us take an example to explain this.

Illustration 9.

Compute the Rank Correlation coefficient from the following data:

Series A	:	115	109	112	87	98	98	120	100	98	118
Series B	:	75	73	85	70	76	65	82	73	68	80

Solution :

Computation of Rank Correlation coefficient from the following data:

Series A	Ranks R_1	Series B	Ranks R_2	D (R_1 - R_2)	D^2
115	8	75	6	2	4
109	6	73	4.5	1.5	2.25
112	7	85	10	3	9
87	1	70	3	2	4
98	3	76	7	4	16
98	3	65	1	2	4
120	10	82	9	1	1
100	5	73	4.5	0.5	0.25
98	3	68	2	1	1
118	9	80	8	1	1
Total N = 10				D_2 = 42.50	

Apply formula to calculate Rank Correlation

$$\rho = 1 \quad \frac{6\ D^{2} + \frac{1}{12}(m^{3}\ m) \quad \frac{1}{12}(m^{3}\ m)}{N^{3}\ N}$$

$$\rho = 1 \quad \frac{6\ 42.50+ \quad \frac{1}{12}(3^{3}\ 3) \quad \frac{1}{12}(2^{3}\ 2)}{10^{3}\quad 10}$$

$$\rho = 1 \quad \frac{6(42.50+2+0.50)}{1000\text{-}10} \quad \rho = 1 \quad \frac{270}{990} \quad 0.727.$$

5. Concurrent Deviation Method

This is the simplest methods of studying correlation. The only thing to be computed under this method is the direction of change of both the variables. The formula is $r_c = \sqrt{\frac{2C\text{-}N}{N}}$

where r_c = Coefficient of concurrent deviations.

C = Number of concurrent deviations.

N = Number of pairs of deviations compared.

The procedure of calculating coefficient of correlation under this method is quite simple as explained.

%4 u Compute the direction of change for both the variables comparing with the proceeding values and put + sign increase and - sign for decrease and 0 for no change.

%4 u Denote these two columns by D_x and D_y .

%4 u Multiply D_x with D_y and determine the value of C which means positive sign only.

%4 u Apply the formula. Take an illustration.

Illustration 10

Calculate the coefficient of correlation by concurrent deviation from the following data:

X	:	100	120	135	135	115	110	120
Y	:	50	40	60	80	80	55	65

Solution :

X	dx	Y	dy	D_xD_y
100		50		
120	+	40	—	—
135	+	60	+	+
135	0	80	+	0
115	—	80	0	0
110	—	55	—	+
120	+	65	+	+
N = 6				C = 3

$$r_C \quad \sqrt{=\frac{2C\text{-}N}{N}} \quad \sqrt{\frac{2\ 3\ 6}{6}} \quad \sqrt{\frac{0}{6}} \quad 0$$

Therefore, the correlation does not exist between the variables.

%4 u Self Assessment Fill in the blanks:

%4 uSpearman's rank correlation is a .... method of computing correlation between two characteristics.

%4 uIn Spearman's rank correlation, various items are assigned.. according to the two characteristics and a correlation is computed between these ranks.

%4 uThe coefficient of correlation obtained on the basis of ranks is called ....

(C) Self check Exercise

%4 uWhat do you mean by Correlation>

%4 uDifferentiate between correlation and causation. 4.8. Summary

Correlation analysis aims to study the extent and nature of relationship between different variables. This can be used to study the behavior of economic and commercial phenomenon because of some sort of relationship amongst them. So, we study how an independent variable affects the dependent variable. The effect of correlation reduce the range of uncertainty in our prediction.

4.9 Glossary:

Correlation analysis: Correlation analysis attempts to determine the 'degree of relationship' between variables.

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

Scatter Diagram : Let the bivariate data be denoted by (Xi, Yi), where i= 1, 2 ...... n. In order to have some idea about the extent of association between variables X and Y, each pair (Xi, Yi), i =1, 2......n, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

Spearman's Rank Correlation: This is a crude method of computing correlation between two characteristics. In this method, various items are assigned ranks according to the two characteristics and a correlation is computed between these ranks.

Univariate Distribution: Distributions relating to a single characteristics are known as univariate Distribution.

Bivariate Distribution : When various units under consideration are with regard to two characteristics, we get a Bivariate Distribution.

Simple correlation: In a correlation analysis, if only two variables are studied it is called simple correlation.

Multiple Correlations: If three or more variables, are studied simultaneously, it is called multiple correlation. For example, when we studt the relationship between the yields of rice with both rainfall and fertilizer together, it is a problem of multiple correlation.

Linear Correlation: In a correlation analysis, if the ratio of change between the two sets of variables is same, then it is called linear correlation.

Zero Correlation : If there is no correlation between variables it is called zero correlation. In other words, if the value of one variable cannot be assogiated with the values of the other variable, it is zero correlation.

4.10 Answers : Self Assessment

1. Univariate Distribution	2. association
3. Correlation Coefficient	4. covariation
5. Correlation Coefficient	6. -1, +1
7. scatter diagram	8. crude
9. Ranks	10. Spearman's rank correlation

%4 u Refer to section 4.3

%4 u Refer to section 4.5

4.11. Terminal Questions:

%4 u Define Correlation. What are the different uses of correlation?

%4 u Explain the conditions under which Rank' correlation and Karl Pearson's correlation is applied.

%4 u Find out Karl Pearson's coefficient of correlation, when following

information is given:-N=5, $\bar{X}$ =10, $\bar{Y}$=20, =100
y= 160, x= 80

%4 u Two judge in a beauty competition rank the 12 entries as follows:

X : 1	2	3	4	5	6	7	8	9	10	11	12
Y: 12	9	6	10	3	5	4	7	8	2	11	1

What degree of agreement is there between the two judges.

4.12 Suggested Readings

Gupta, S.P., Statistical Methods, Sultan Chand & Sons, New Delhi.

Gupta, C.C. Fundamentals of Statistics.

Croxton and Cowden, Applied General Statistics.

Hooda, R.P. Statistical for Business and Economics, MacMillan, New Delhi.

*****

Lesson-5
Regressing Analysis

Structure

5.1. Learning Objectives:

After studying the lesson, you should be able to understand:

%4 u What is Regression Analysis?
%4 u Points of Difference between correlation and Regression.
%4 u Methods of Regression Analysis.
%4 u Regression Equations.
%4 u Regression Coefficient and its properties.

5.2. Introduction

The principal of correlation establishes the degree and direction of relationship between two or more variables. But we may be interested in estimating the value of an unknown variable on the basis of a known variable. If we know the indices of supply of money and price-level, we can find out the degree and direction of relationship between these indices with the help of correlation technique. But the regression technique helps us in knowing what the general price-level would be assuming a fixed supply of money. Similarly if we know that the price and demand of a commodity are correlated we can find out the demand for that commodity for a fixed price. Hence, the statistical tool with the help of which we are in a position to estimate or predict the unknown values of one variable from known values of another variable is called regression. The meaning of the term “Regression” is the act of returning or going back. This term was first used by Sir Francis Gallon in 1977 when he was studying the relationship between the height of fathers and sons. His study pointed out a very interesting relationship. All tall fathers tend to have tall sons and all short fathers short sons but the average height of the sons of a group of tall fathers is less than that of the fathers and the average height of the sons of a group of short fathers is greater than that of the fathers. The line describing the tendency to regress or going back was called a “Regression Line”. These days the modern writers have started to use the term estimating line instead of regression line because the expression estimating line is more clear in character. According to Morris Myers Blair, regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

5.3. Regression Analysis

Regression analysis is a branch of statistical theory which is widely used in all the scientific disciplines. It is a basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. The uses of regression are not confined to economic and business fields only. Its applications are extended to almost all the natural, physical and social sciences. The tool of regression can be extended to three or more variables but we shall confine ourselves to the problems of two variables in this lesson which is called simple regression.

Regression analysis is of great practical use even more than the correlation analysis. Some of the uses of the regression analysis are:

%4 u Regression Analysis helps in establishing a functional relationship between two or more variables. Once this is established it can be used for various analytic purposes.

%4 u With the use of electronic machines and computers the medium of calculation of regression equations particularly expressing multiple and non-linear relation has been reduced a great deal.

%4 u Since most of the problems of economic analysis are based on cause and effect relationship, the regression analysis is a highly valuable tool in economic and business research.

%4 u The regression analysis is very useful for prediction purposes. Once a functional relationship is known, the value of the dependent variable can be predicted from the given value of the independent variables.

5.4. Difference between Correlation and Regression

The two techniques are directed towards a common purpose of establishing the degree and direction of relationship between two or more variables but the methods of doing so are different. The choice of one or the other will depend on the purpose. For example, if the purpose is to know the degree and direction of relationship, correlation is alright but if the purpose is prediction for nature of a dependent variable with the substitution of one or more independent variables, the regression analysis shall be more helpful. The point of difference are discussed below:

%4 u Degree and Nature of Relationship: The correlation coefficient is a measure of degree of co-variability between two variables whereas the regression analysis is used to study the nature of relationship between the variables so that we can predict the value of one on the basis of another. The reliance on the estimates or predictions depends upon the closeness of relationship between the variables.

%4 u Cause and Effect Relationship: The cause and effect relationship is explained by regression analysis. Correlation is only a tool to ascertain the degree of relationship between two variables and we can not say that one variable is the cause and the other the effect. A high degree of correlation between price and demand for a commodity or at a particular point of time may not suggest which is the cause and which is the effect. However, in regression analysis cause and effect relationship is clearly expresses-one variable is taken as dependent and the other as independent.

Conventionally, the variable which is the basis of prediction is called independent variable and the variable that is to be predicted is called dependent variable. The independent variable is denoted by X and the dependent variable by Y.

5.5. Principle of Least Squares

Regression refers to an average of relationship between a dependent variable with one or more independent variables. Such relationship is generally expressed by a line of regression drawn by the method of the “Least Squares”. This line of regression can be drawn graphically or derived algebraically in the form of an equation of line by any method explained in this lesson. Before we discuss the various methods let us understand the meaning of the least squares assumption. According to TOM CARS, before the equation of the least line can be determined some criteria must be established as to what conditions the best line should satisfy. The condition usually stipulated in regression analysis is that the sum of the squares of the deviations of the observed ‘Y’ values from the fitted line shall be minimum. This is known as the least squares or minimum squared error criterion.

A line fitted by the method of least squares is the line of best fit. It is the line which satisfies the following conditions:

The algebraic sum of deviations above the line and below the line are equal to zero. We can state:

$(x\ x_c) = 0$

$(y\ y_c) = 0$

Where x and y are the values derived with the help of regression analysis.

The sum of the squares of all these deviations is less than the sum of the squares of deviations from any other line, we can say

$(y\ \ y_c)_2$ is smaller than $(y - A)_2$

and

$(x - x_c)^2$ is smaller than $(x - A)^2$

Where A is a corresponding value on any other straight line.

%4 u The line of regression (best fit) intersect at the mean value of the variables, say at X

and Y . – –

%4 u When the data represent a sample from a larger population, the least square line is the best estimate of the population line.

5.6. Methods of Regression Analysis

We can study regression by two methods:

%4 u Graphic method (regression lines)

%4 u Algebraic method (regression

equations) We shall discuss these methods

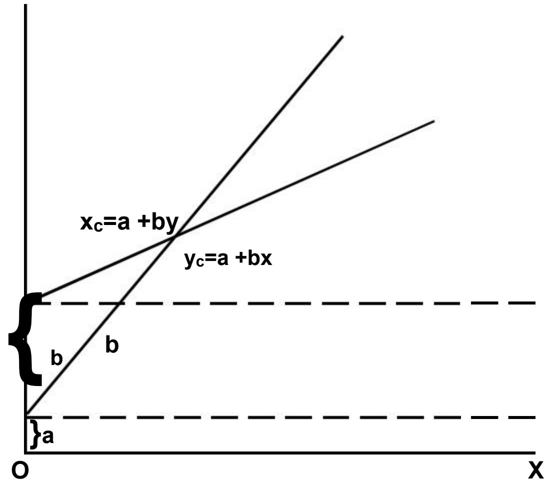
individually.

%4 u Graphic Method: Under this method the various points are plotted on a graph paper representing various pairs of values. These points give a picture on a scatter diagram with several points spread over. A regression line may be drawn in between these point either by free hand or by a scale in such a way that the squares of the vertical or horizontal distances between the points and the line of regression so drawn is the least. It should be drawn faithfully as the line of best fit leaving equal number of points on both sides in such a manner that the sum of the squares of the distance is the least. However, to ensure this is rather difficult and therefore, the method only renders a rough estimate which can not be completely free from subjectively of person drawing it. Such a line can be a straight line or a curved line depending upon the scatter of points and relationship sought to be established. A nonlinear free hand curve will have more element of subjectivity and normally a straight line is drawn. This can be best explained with the help of a below given example:

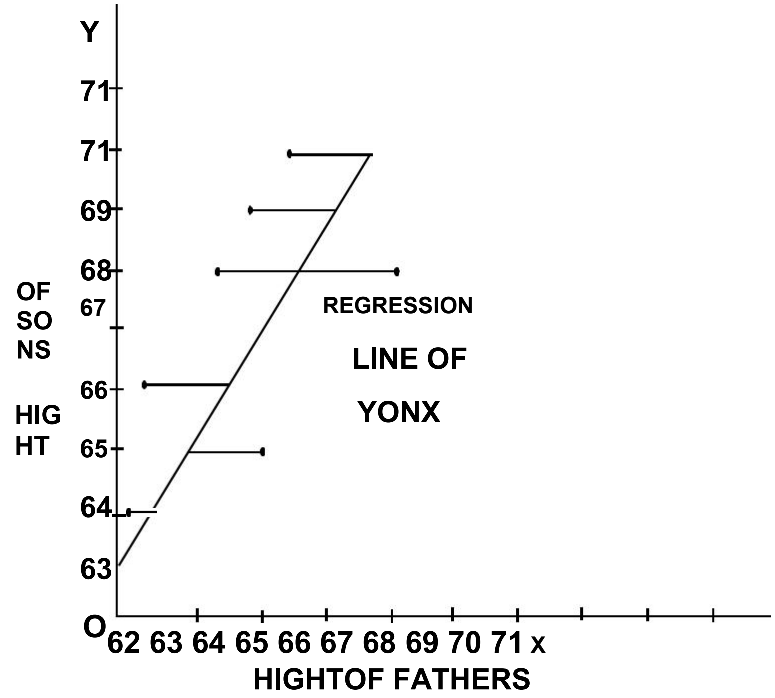
Example 1.

Height of Father (Inches)	Height of Sons (Inches)
65	68
63	66
67	68
64	65
68	69
62	66
70	68
66	65
68	71
67	67
69	68
71	70

The following diagram above the height of fathers on x-axis and the height of sons on y-axis. The line of regression called the regression of y on x is drawn in between the scatter dots.

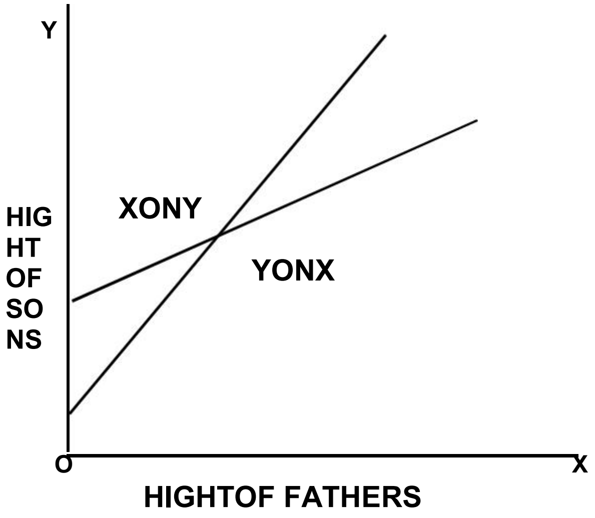


Another line of regression called the regression line of x on y is drawn amongst the same set of scatter dots in such a way that the squares of the horizontal distances between dots are minimized.



It is clear that the position of the regression line of x on y is not exactly like that of the regression line of y on x.

In the following figure both the regression of y on x and x on y are exhibited.



When there is either perfect positive or perfect negative correction between the two variables the two regression lines will co-inside we will have only one line. The farther the two regression lines from each other, the lesser is the degree of correlation and vice-versa. If the variables are independent, correlation is zero and the lines of regression will be at right angles. It should be noted that the regression lines cut each other at the point of average of x and y, i.e.. if from the point where both the regression lines cut each other a perpendicular is drawn on the x-axis, we will get the mean value of x and if from that point a horizontal line is drawn on the y-axis we will get the mean of y.

%4 u Algebraic Method: The algebraic method for simple linear regression can be broadly divided into two class:

(i) x on y is used to describe the variations in the values of x for given changes in y.

(ii) y on x is used to describe the variations in the values of y for given changes in x.

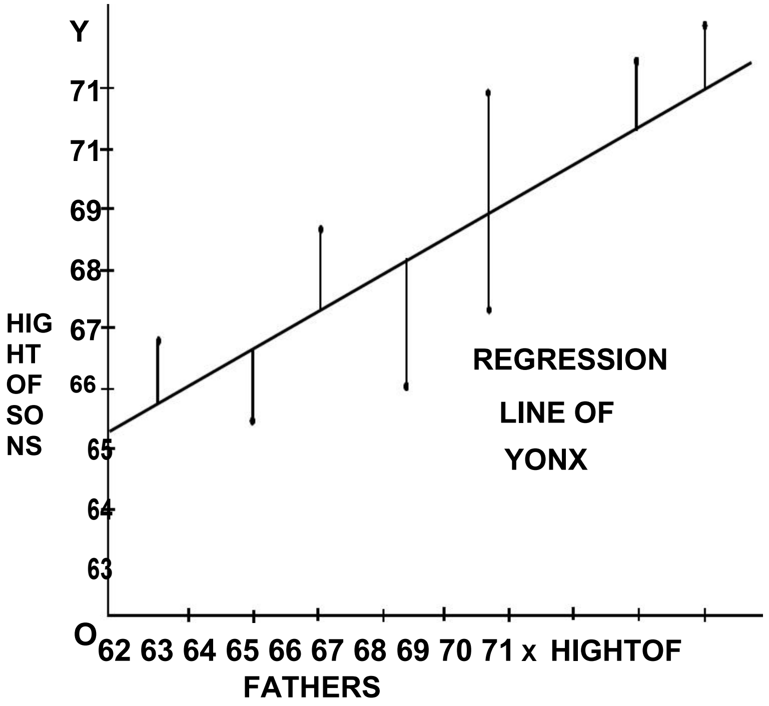
%4 u Regression Equations: These equations are known as estimating equations. Regression equations are algebraic expressions of the regression lines. As there are two regression lines, there are two regression equations:

%4 u x on y is used to describe the variations in the values of x for given changes in y.

%4 u y on x is used to describe the variations in the values of y for given changes in x. The regression equations of y on x is expressed as $y_c = a + bx$

The regression equations of x on y is expressed as $x_c = a + by$

In these equations a and b are constants which determine the positions of the line completely. These constants are called the parameters of the line. If the value of any of these parameters is changed, another line is dermined.



The parameter a refers to the intercept of the line and b to the slope of the line. The symbol y_c and x_c refers to the values of y computed and the value of x computed on the basis of independent variable in both the cases. If the values of both the parameters are obtained. The line is completely determined. The values of these two parameters a and b can be obtained by the method of least squares. With a little algebra and differential calculus it can be shown that the following two equations, is solved simultaneously will give values of the parameters a and b such that the least squares requirements is fulfilled;

For regression equation $y_c = a + bx$

$$y = Na + b x$$
$$xy = a x + b x_2$$

For regression equation $x_c = a + by$

$$x = Na + b y$$
$$xy = a y + b y_2$$

These equations are usually called normal equations. In the equations x, y, xy, x_2, y_2 indicate totals which are computed from the observed pairs of values of two variables x and y to which the least squares estimating line is to be fitted and N is the number of observed pairs of values. Let us explain it by an example.

Example 2. From the following data obtain the two regression equations:

x:	6	2	10	4	8
y:	9	11	5	8	7

Solution :

Computation of Regression Equations

X	y	xy	x^2	y^2
6	9	54	36	81
2	11	22	5	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
x = 30	y = 40	xy = 214	x_2= 220	y_2 = 340

Regression line of y on x is expressed by the equation

of the form $y_c = a + bx$

To determine the values of a and b the following two normal equations

are to be solved $y = Na + b x$

$xy = a x + b x_2$

Substituting the values, we get

$40 = 5a + 30b$ ...(i)

$214 = 30a + 220b$ ...(ii)

Multiplying equation (i) by 6, we get

$240 = 30a+180b$ ...(iii)

$214 = 30a + 220b$ ...(iv)

Deduct equation (iv) and (iii)

$- 40b = + 26$

$b = - 0.65$

Substitute the value of b in equation (i)

$40 = 5a + 30 (- 0.65)$

$5a = 40 + 19.5$

$a - 11.9$

Put the value of a and b in the equation The regression of y on x is

$y = 11.9-0.65x$

Regression line of x on y is expensed by the equation

$xc = a + by$

The corresponding normal equations are

$x = Na + b y$

$xy = a y + b y_2$

Substitute the values

$30 = 5a + 40b$ ...(i)

$214 = 40a + 340b$ ...(ii)

Multiple equation (i) by 8

$240 = 40a + 320b$ ...(iii)

$214 = 40a + 340b$ ...(iv)

Deduct equation (iv) from (iii)

-20b = 26

b = - 1.3

Substitute the value of b in equation (i)

%4 u= 5a +

40(-1.3) 5a = 30

+ 52

a = 16.4

Put the values of a and b in the equation. The regression line of x on y is X = 16.4 - 1.3y.

5.8. Regression Coefficient: In the regression equation ‘b1 is the regression coefficient which indicates the degree and direction of change in the dependent variable with respect to a change in the independent variable. In the two regression equations :

X = a + bxy

Y = a + byx

Where bxy and byx are known as the regression coefficient of the two equations. There coefficient can be obtained independently without using simultaneous normal equations with these formulae :

Regression coefficient of x on y is

$$bxy = r \frac{x}{y}$$

xy$\underline{x}$x

bxy =2 ——— ———

x y y y

$$bxy = \frac{xy}{y^2}$$ where x = $\bar{x}$ and y - $\bar{y}$

Regressing Coefficient of y on x is Y = a + byx

$$byx = r \frac{y}{x}$$

$$byx = \frac{xy}{x\ y} \quad \frac{x}{y} = \frac{xy}{x^{\ 2}}$$ where x = x-$\bar{x}$ and y-$\bar{y}$

Example 3. Calculate the regression coefficients from data given below :

	Series x	Series y	
Average	25	22	
Standard deviation	4	5	r = 0.8

The coefficient of regression of x on y is $bxy = r \frac{x}{y}$

$$= 0.8 \ \frac{5}{4} \quad .64$$ the coefficient of regression of you x is

$$bxy = r \frac{x}{y} = 0.8 \quad 5^{\bar{5}} \ 1.00$$

5.9. Properties of Regression Coefficients

(i) The coefficient of correlation is the geometric mean of the two regression coefficients.

$$r = \sqrt{bxy \quad byx}$$

%4 u Both the regression coefficients are either positive or negative. It means that they always have identical sign.

%4 u The Coefficients of correlation and the regression coefficients will have identical sign.

%4 u If one of the regression coefficients is more than unity, the other must be less unity because the value of coefficient of correlation can not exceed one (r = ±1).

%4 u Regression coefficients are independent of the change in the origin but not of the scale.

%4 u The average of regression coefficients is always less than correlation coefficient. We can compute the regression equation with the help of regression coefficients by the following equations:

1. Regression equation x on y

$$X - \underline{X} = \underline{x}\,(Y-\underline{Y})$$

y

where $\overline{X}$ is the mean of X se$\overline{\text{ri}}$es Y is the mean of Y series

$r\,\frac{\underline{x}}{y}$ is known as the regression coefficient of x on y.

2. Regression equation of y on x

$$Y - \overline{Y} = r\,\frac{x}{y}\,(X-\overline{X})$$

(A)Self Assessment Fill in the blanks:

1. Study of .... meant to determine the most suitable form of the relationship between the variables given that they are correlated

2. Coefficient of correlation is measure of the degree of .... of the variables.

3. The regression equations are useful for predicting the value of .... variable for given value of the independent variable.

4. The nature of a regression equation is .... the nature of a mathematical equation.

%4 uThe term regression was first introduced by Sir Francis Galton in.................

%4 uIf X is independent variable then we can estimate the average values of Y for a given value of X. The relation used for such estimation is called regression of

%4 u If Y is used for estimating the average values of X, the relation will be called regression of ....

8. For a bivariate data, there will always be.............. of regression.

We can explain this by taking an example:

Example 4. Calculate the following from the below given data:

%4 u the two regression equations,

%4 u the coefficients of correlation and

%4 u the most likely marks in Statistical when the marks in Economics are 30.

Marks in Economics :	25	28	35	32	31	36	29	38	34	32
Marks in Statistics :	43	46	49	41	36	32	31	30	33	39

Solution :

Calculation of Regration Equations and Correlation Coefficient

Marks in Economics X	$(X-\bar{X})$ x	x^2	Marks in Statistics Y	$(Y-\bar{Y})$ y	y^2	xy
25	-7	49	43	+ 5	25	-35
28	-4	16	46	+ 8	64	-32
35	+ 3	9	49	+ 11	121	+ 33
32	0	0	41	+ 3	9	0
31	- 1	1	36	-2	4	+ 2
36	+ 4	16	32	-6	36	-24
29	-3	9	31	-7	49	+ 21
38	+ 6	36	30	-8	64	-48
34	+ 2	4	33	-5	25	- 10
32	0	0	32	+ 1	1	0
X=320	x = 0	x^2 = 140	Y=380	y = 0	y^2 = 398	xy=-93

Regression equation of X on Y $\bar{X} - \bar{Y}$ = bxy (Y - $\bar{Y}$) :

$$bxy = \frac{xy}{y^2}$$

$$bxy = 398 \frac{93}{} \quad 0.234$$

$$\bar{X} \quad N \underline{X} \quad \frac{320}{10} \; 32$$

$$\bar{Y} \quad N \underline{X} \quad \frac{380}{10} \; 38$$

Substituting the values

X-32 = - 0.234 (Y—38)

X-32 = - 0.234 Y + 8.892 or X = 40.892 -0.234 Y

Regression equation of Y on X

$Y - \bar{Y} = byx (X - \bar{X})$

$$byx = \frac{xy}{x^2} \quad \frac{93}{140} \; 0.664$$

$\bar{}$%4 u = 32

$\bar{}$%4 u = 38.

b = - 0.664

Y = 38. b = - 0.664 (x - 38) = 0.664 X +21.248 y = 59.248-0.664 X.

(b) Correlation Coefficient

$$= \sqrt{bxy \quad byx}$$

$$\%4 \quad u \quad \sqrt{0.234 \quad -0.664} \quad = -0.394.$$

Since both the regression coefficients are negative, value of r must also be negative.

%4 u Likely marks in statistics when marks in Economics are 30. y = -0.664 X + 59.248

where X=30

y = (-0.664 30) + 59.248 = 39.328 = 39.328 or 39.

Example 5. The following scores were worked out from a test in Mathematics and English in an annual examination:

	Scores in	
	Mathematics (x)	English (y)
Mean	39.5	47.5
Mean		
Standard deviation	10.8	16.8

r = + 0.42

Find both the regression equations. Using these regression estimate find the value of Y for X = 50 and the value of X for Y = 30.

Solution.

Regression of X on :

$X - \underline{X} = \underline{x} \ (Y - \underline{Y})$

y

where $\bar{}$Y =$\bar{}$47.5, X = 39.5

r = 0.42, = 10.8, = 16.8

‘x y

Put these values in the equation

$$X - 39.5 = 0.42 \ \frac{10.8}{16.8} = (Y - 47.5)$$

or X = 0.27 (Y - 47.5) = 0.27 Y-12.82

X = 0.27 Y- 12.82 + 39.5 = 0.27 Y +26.68

When y = 30

The value of X is (0.27 x 30 + 26.68) - 34.78 Regression equation of Y on X :

%4 u _ $\underline{x}$ X - X

Y- Ұ

y

where

= 39.5, Y = 47.5

r = 0.42, $_y$ =16.8, $_x$ = 10.8.

Substitute these values

$$Y - 47.5 = 0.42 \ \frac{16.8}{10.8} \ (X - 39.5)$$

= 0.653 (X - 39.5) = 0.653 X - 25.79

or = 0.653 X-25.79+ 47.5 = 0.653 X + 21.71

When X = 50

The value of y is (0.653 50 + 21.71)

y = 32.65 + 21.71 = 54.36

Thus the regression equations are

$x = 0.27\ y + 26.68$

$y = 0.653\ x + 21.71$

Value of x when y = 30 is 34.78

Value of y when x = 50 is 54.36

When actual mean of both the variables x and y come out to be in tractions, the deviations from actual means create a problem and it is advisable to take deviations from the assumed mean. Thus when deviations are taken from assumed means, the value of bxy and byx a given by

$$bxy = \frac{dxdy - \frac{(\ dx)\ (\ dy)}{N}}{dx^2 \frac{(\ dx)^2}{N}}$$

where dx = (X - A)
and dy = (Y - A)

The regression equation is

$X - \overline{X} = bxy\ (Y - \overline{Y})$ Similarly the regression equation of y on x is

$Y - \overline{Y} = byx\ (X - \overline{X})$

$$byx = \frac{dxdy - \frac{dxdy}{N}}{dx^2 \quad \frac{(\ dx)^2}{N}}$$

Let us take an example to explain this.

Example 6. You are given the data relating to purchases and sales. Compute the two regression equations by methods of least squares and estimate the likely sales when the purchase equal 100.

Purchases:	62	72	98	76	81	56	76	92	88	49
Sales :	112	124	131	117	132	96	120	136	97	85

Solution :

Calculations of Regression Equations.

Purchases	(X -76)		Sales	(Y-120)		
X	dx	dx_2	Y	dy	dy_2	dxdy
62	-14	196	112	-8	64	+ 112
72	- 4	16	124	+ 4	16	16
98	+ 22	484	131	+ 11	121	+ 242
76	0	0	117	-3	9	0
81	+ 5	25	132	+ 12	144	+ 60
56	-20	400	96	-24	576	+ 480
76	0	0	120	0	0	0
92	+ 16	256	136	+ 16	256	+ 256
88	+ 12	144	97	-23	529	-276
49	-27	729	85	-35	1225	+ 945
	dx = -10	dx_2 - 2250		dy = - 50	dy_2 = 2940	dxdy = 1803

$$\bar{X} = A + \frac{dx}{N}\ 76\quad \frac{10}{10}\ 75\ ,\qquad \bar{Y} = A\ \frac{dy}{N}\ 120\quad \frac{50}{10}\ 115.$$

Regression Coefficients :
X on Y

$$bxy = \frac{dxdy - \frac{(\ dx)\ (\ dy)}{N}}{dx^2\quad \frac{(\ dx)^2}{N}}$$

$$= \frac{1803 - \frac{(10)\ (-50)}{10}}{2940\quad \frac{(50)^2}{10}}\quad \frac{1753}{2690}\ 0.652.$$

$$\text{Y on X bxy} = \frac{dxdy - \frac{(\ dx)\ (\ dy)}{N}}{dx^2\quad \frac{(\ dx)^2}{N}}$$

$$= \frac{1803 - \frac{(10)\ (-50)}{10}}{2950\quad \frac{(10)^2}{10}}\quad \frac{1753}{2240}\ 0.78.$$

Regression equations : X on Y

$X\quad \bar{X} = bxy\ (\ Y\quad \bar{Y}\ )$

Substitute the values

X-75 = 0.652 (Y-115) = 0.652 Y - 74.98

X= 0.652 Y +0.02

whenX = 100

Y = 0.78 100 + 56.5 = 134.5

Regression equations of Y on X :

$Y\ \bar{Y} = byx\ (\ X\ X\ )\ \bar{Y}$-115 = 0.78 (X-75) = 0.78 X-58.5 = 0.78 X +56.5

Standard Error of an Estimate

Standard error of an estimate is the measure of the spread of observed values from the estimated ones, expressed by regression line or equation. The concept of standard error of an estimate is analogous to the standard which measures the variation or scatter of individual items about the arithmetic mean. Therefore, like the standard deviation which is the average of square of deviations about the arithmetic mean, the standard error of an estimate is the average of the square of deviations between the actual or the observed values and the estimated values based on the regression equation. It can also be expressed as the root of the measure of unexplained variations divided by N-2:

$$Sxy\ \sqrt{\frac{\text{Unexplained variation}}{N-2}}$$

$$Syx\ \sqrt{\frac{(Y\quad \bar{Y}_c)^2}{N-2}}$$

$$\text{And Sxy} \quad \sqrt{\frac{(X \quad \overline{X}_c)^2}{N-2}}$$

Where Syx refers to standard error of estimate of Y values on X values.

Sxy refers to standard error or estimate of X values on Y value.

Y_c and X_c are the estimated values of Y and X variable by means of their regressions equations respectively. N - 2 is used for getting an unbiased estimate of standard error. The usual explanation given for this division by N %4 u 2 is that the two constants a and b were calculated on the basis of original data and we lose two degress of freedom. By degrees of freedom we mean the number of classes to which values can as a signed at will without violating any of the restrictions imposed. However a simplex method of computing Syx and Sxy is to use formulae :

$$\text{Syx} \quad \sqrt{\frac{Y^2 \quad a \;\; y - b \;\; XY}{N-2}}$$

$$\text{and} \qquad \text{Sxy} \quad \sqrt{\frac{X^2 \quad a \;\; X - b \;\; X}{N-2}}$$

The standard error of estimate measures the accuracy of the estimated figures. The smaller the values of standard error of estimate, the closer will be the dots to the regression line and the better the estimates based on the equation for this line. If standard error of estimate is zero, then there is no variation about the line and the correlation will be perfect. Thus with the help of standard error of estimate it is possible for us to ascertain how good and representive the regression line is as a description of the average relationship between two series.

Example 7. Given the following data :

X	6	2	10	4	8
Y	9	11	5	8	7

And two regression equations Y = 11.09-0.065 andX= 16.4- 1.3 Y. Calculate the standard error of estimate i.e.. Syx and Sxy.

Solution :

We can calculate X_c and Y_c values from these regression equations.

X	Y	Y_c	X_c	$(Y-Y_c)^2$	$(X-X_X)^2$
6	9	8.0	4.7	1.00	1.69
2	11	10.6	2.1	0.16	0.01
10	5	5.4	9.9	0.16	0.01
4	8	9.3	6.0	1.69	4.00
8	7	6.7	7.3	0.09	0.49
X-30	Y=40	$Y_c = 40$	$Y_c = 30$	$(Y- Y_c)^2 = 3.1$	$(X-X)^2_c = 6.20$

Thus we can calculate Syx and Sxy from the above calculated values.

$$\text{Syx} \quad \sqrt{\frac{(Y \quad \overline{Y}_c)^2}{}} \quad \sqrt{\frac{3.1}{}} \quad \sqrt{1.03 \;\; 1.01}$$

$$\text{Sxy} \quad \sqrt{\frac{(Y \quad \overline{Y}_c)^2}{N-2}} \quad \sqrt{\frac{6.2}{5-2}} \quad \sqrt{}\; 2.07 \;\; 1.01$$

(B) Setf Assessment

State whether the following statements are true or false:

%4 u Least square root method is one of the most popular methods of fitting a mathematical trend.

%4 u The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, are minimized

%4 u The general form of linear trend equation is Yt = a + bt.

%4 u The general form of an exponential trend is Y = a.bt

%4 u The results of the method of least squares are most satisfactory.

%4 u Least square method can be used to fit growth curves.

(C) Self check exercise

%4 u What do you mean by Regression?

%4 u What are the utility of Regression?

%4 u Differentiate between Regression

and Correlation. 5.11. Summary

Regression Analysis provides estimates of the value of the Dependent variable from the value of Independent variable. This analysis also helps in measuring the errors involved in using regression line as the basis of estimation. So Regression analysis attempts to establish the nature of the relationship between variables that is to study the functional relationship between variables and provide a mechanism for prediction or forecasting.

5.12 Glossary:

Regression : The statistical method which helps us to estimate or predict the unknown value of one variable from the known value of the related variable is called regression.

Regression equation : Regression equations are the algebraic formulation of regression lines. The regression equations may be regarded as expressions for estimating from a given value of one variable the average corresponding value of the other.

Exponential trend: The general form of an exponential trend is Y = a.bt, where a and b are constants.

Least square methods: This is one of the most popular methods of fitting a mathematical trend. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, are minimized.

Line of Regression Y on X: The general form of the line of regression of Y on X is YCi=a + bXi, where YCi denotes the average or predicted or calculated value of Y for a given value of X =Xi. This line has two constants, a and b.

Line of Regression of X on Y: The general form of the line of regression of X on Y is XCi=c + dYi, where XCi denotes the predicted or calculated or estimated value of X for a given value of Y= Yi and c and d are constants. d is known as the regression coefficient of regression of X on Y

Linear Trend: The linear trend equation is given by relation Yt = a + bt. where t denotes time period such as year, month, day, etc., and a, b are the constants.

5.13 Answers: Self Assessment

1. 'Regression'
2. linear association
3. dependent:
4. different from.
5.1877
6.YonX
7. Xony
8. Two lines
9. False
10. True
11. True
12. True
13. True
14. False
15. Refer to section 5.3
16. Refer to section 5.3
17. Refer to section 5.4

5.14 Terminal Questions

%4 u What is the difference between Correlation and Regression?

%4 u What do you mean by regression Coefficients? What are the different characteristics of regression coefficients?

%4 u Explain the concept of standard Error of Estimate.

%4 u From the following data find the two regression equations.

X	1	2	3	4	5
Y	2	3	5	4	6

Find the most probable value of Y when X is 2.5.

5.15. Suggested Readings:

Gupta S.P. : Statistical Methods

Gupta S.C. : Fundamentals of Statistics

Corxton and Cowden : Applied General Statistics

Chou : Statistical Analysis

Kazmier Leonard J : Business Statistics.

*****

Lesson -6
Index Numbers

Structure

6.1. Learning Objectives

After studying the lessons, you should be able to understand:

%4 u The concept of Index Numbers

%4 u What are the different problems involved in the preparation of index Numbers.

%4 u What are the different Methods of Construction of Index Numbers.

6.2. Introduction:

Economic activities have constant tendency to change. Prices of commodities which are the total result of number of economic activities also have a tendency to fluctuate. The problem of change in prices is very important. But it is not very simple to study this problem and derive conclusion because prices of different commodities change by different degress. Hence, there is a great need for a device which can smoothen the irregularities in the prices to obtain a conclusion. This need is satisfied by Index Number which makes use of percentage and averages for achieving the desired objective. Index Number is a device for comparing the general level of the magnitude of a group of distinct but related variables in two or more situations. Index Numbers are used to feel the pulse of the economy and they reveal the inflationary or deflationary tendencies. In reality, Index Numbers are described as barometers of economic activity because if one wants to have an idea as to what is happening in an economy, he should check the important indicates like the index number of industrial production, agricultural production, business activity etc.

6.3. Defination:

The various definitions of Index Number can be discussed under three heads :

%4 u Measure of change

%4 u Device to measure change

%4 u A series representing the process of change.

According to Maslow, it is a numerical value characterizing the change in complex economic phenomenon over a period of time.

Spiegal explains an index number is a statistical measure designed to show changes in variable or a group of related variables with respect tc time, geographical location or other characterisitics.

Gregory and Ward describes it as a measure over time designed to show average change in the price, quality or value of a group of items.

Croxton and Cowden says that Index numbers are devices for measuring differences in the magnitude of a group of related variables.

R.L. Bowley describes Index Numbers as a series which reflects in its trends and fluctuations the movements of some quantity to which it is related.

Blair puts Index Numbers are specialized kinds of an average.

Index Number have the following features:

%4 u Index Numbers are specialized average which are capable of being expressed in percentage,

%4 u Index Numbers measures the changes in the level of a given phenomenon.

%4 u Index numbers measures the effect of changes over a period of time.

Index Numbers are indispensable tools of economic and business analysis. Their significance can be appreciated by these points:

%4 u Index Number helps in measuring relative changes in a set of items.

%4 u Index numbers provide a good basis of comparison because they are expressed in abstract unit distinct from the unit of element.

%4 u Index numbers help in framing suitable policies for business and economic activities.

%4 u Index numbers help in measuring the general trend of the phenomenon.

%4 u Index numbers are used in deflating. They are used to adjust the original data for price changes or to adjust wages for cost of living changes.

%4 u The usefulness of index number has grown a great deal because of the method of splicing whereby the index prepared on any one base can be adjusted with reference to any other base.

%4 u As a measure of average change in a group of elements the index numbers can be used for forecasting future events. Whereas a trend line gives an average rate of change in a single phenomenon, it indicates the trend for a group phenomenon.

%4 u It is helpful in a study of comparative purchasing power of money in different countries of the world.

%4 u Index numbers of business activity throw light on the economic progress that various countries have made.

%4 u Problem in the Construction of Index Numbers

The construction of Index Number involves the consideration of the following important problems:

%4 u The Purpose of Index: Before constructing an Index number, it is necessary to define precisely the purpose for which they are to be constructed. There is no Index numbers which can fulfill all the purposes. Index numbers are specialized tools which are more efficient and useful when properly used. If the purpose is not clear, the data used may be unsuitable and the indices obtained may be misleading. If it is desired to construct a Cost of Living Index number of Labour class, then only those item will be included, which are required by the labour class.

%4 u Selection of the items or ‘Regimen’: The list of commodities included in the Index numbers is called the ‘Regimen’. Because it may not be possible to include all the items, it becomes necessary to decide what items are to be included. Only those items should be elected which are representative of the data, e.g., in a consumer price Index for working class, items like scooters, cars, refrigerators, cosmetics, etc., find no place. There is no hard and fast rule regarding the inclusion of number of commodities while constructing Index Numbers. The number of commodities should be such as to permit the influence of the inertia of large numbers. At the same time the numbers should not be so large as to make the work of computation uneconomical and even difficult. The number of commodities should therefore be i.e., reasonable. In simple words the following broad factors must be duly considered:

(i) The items should be representative.

(ii) The items should be of a standard quality. (iii)
Non-tangible items should be excluded. (iv) The
items should be reasonable in number.

%4 u Price Quotation: It is neither possible nor necessary to collect the prices of the commodities from all the markets in the country where it is dealt with, we should take a sample of the markets. Selection must be made of the representative places and persons. These places should be those which are well known for trading with the commodity.

(a) Money Prices: In which prices are quoted per unit of commodity, e.g., what price at Rs.50 per quintal; and

(b) Quantity Prices: In which prices are quoted per unit of commodity, e.g., at 2 kilogram a rupee. It is necessary to select a reliable agency from where price quotations have to be obtained.

Selection of the Base : In the construction of Index Numbers, the selection of the base period is very important step. The base period serves as a reference period and the prices for a given period are expressed as percentage of those for the base year, it is therefore necessary that :

(i) the base period should be normal and (ii) it should not be too far in the past.

There are two methods by which base period can be selected (1) Fixed base method and (2) Chain base method.

Fixed base Method: According to this any year is taken as a base. Prices during the year are taken equal to 100 and the prices of other years are shown as percentages of those prices of the base year. Thus if indices for 1978, 79,

%4 uand 81 are calculated with 1977 as base year, such indices will be called as fixed base indices.

Chain base Method: According to this method, relatives of each year are calculated on the basis of the prices of the preceding year. The Chain base Index Numbers are called as Link Relative e.g., if index numbers are constructed for 1977, 78, 79, 80 and 81 then for 1978, 1977 will be the base and for 1979, 1978, will be the base and so on.

The choice of an average: An Index Number is a technique of 'averaging' all the changes in the group of series over a period of time, the main problem is to select an average which may be able to summaries the change in the component series adequately. Median, Mode and Harmonic Mean are never used in the construction of Index Numbers. A choice has to be made between the Arithmetic Mean and the Geometric Mean. Merits and demerits of the two are then to be compared. Theoretically G.M. is superior to be A.M. in many respects but due to difficultly in its computation, it is not widely used for this purpose.

Selection of appropriate weights: The term weight refers to the relative importance of the different items in the construction of index numbers. All items are not of equal importance and hence it is necessary to find out some suitable methods by which the varying importance of the different items is taken into account. The system of weighing depends upon the purpose of index numbers, but they ought to reflect the relative importance of the commodities in the regimen. The system may be either arbitrary or rational. The weightage may be according to either:

(1) the value or quantity produced, or

(2) the value or quantity consumed, or

(3) the value or quantity sold or put to sale. There are two methods of assigning weights.

(i) implicit and (ii) Explicit.

Implicit: In this method, the commodity to which greater importance has to be given is repeated a number of times i.e. a number of varieties of such commodities are included in the index numbers are separate items.

Implicit: In this case, the weights are explicitly assigned to commodities. Only one kind of a commodity is included in the construction of Index Numbers but its price relative is multiplied by the figure of weights assigned to it. There has to be some logic in assigning such type of weights.

(A)Self Assessment Fill in the blanks:

%4 uAn .... is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations.

%4 u Index number is often used to average a .... expressed in different units for different items of a group.

%4 uPrice index can be used to determine the .... and .... of average change in the prices for the group.

%4 uThe year from which comparisons are made is called the .... year.

%4 u The year under consideration for which the comparisons are to be computed is called the .... year.

6.5. Methods of Construction of Index Numbers

The index number for this purpose is divided into two categories:

%4 u Un-weighted Indices; and

%4 u Weighted Indices.

Each one of these types may further be sub-divided under two heads

%4 u Simple aggregative; and

%4 u Average of price relatives. Un-weighted Index Numbers

%4 u Simple aggregative method: Under this method the total of the current year prices for various commodities is divided by the total of the base year and the quotient is multiplied by 100.

Symbolically,

$$P_{01} \quad \frac{p_1}{p_2} \quad 100$$

where 'P' represents the Price Index, 'p'—prices, 'I' current year and '0' base year.

Illustration 1. From the following data construct the index for 1981 taking 1980 as base year.

Commodity	Price in 1980 (Rs.)	Prices in 1981 (Rs.)
A	30	30
B	35	50
C	45	75
D	45	70
E	25	40

Solution: Construction of Price Index

Commodity	Price in 1980	Prices in 1981
	(P_0)	(P_1)
A	30	30
B	35	50
C	45	75
D	45	70
E	25	40
	p_0=180	p_1 = 265

Price Index for 1981 with 1980 as base

$$\frac{\text{sum of prices in 1981}}{\text{sum of prices in 1980}} \quad 100$$

Symbolically,

where $P_{01} \quad \frac{p_1}{p_0} \quad 100$

i.e. $P_{01} \quad \frac{265}{180} \quad 100$

OR P_{01} = 14.7.2 Ans.

Hence there is an increase of 47.2% in prices of commodities during the year 1981 as compared to 1980.

%4 u Average of Price Relative Method: Under this method, calculate first the price relative for the various items included in the index and then average the price relative by using, say, of the measures of the central value, i.e.. A.M.; the median; the mode; the Geometric mean or the Harmonic mean. The following are the formula (with usual symbols) when.

(1) Arithmetic Mean is used $P_{10} = \dfrac{\frac{p_1}{p_0} \; 100}{N}$

(2) Geometric Mean is used P_{01} = antilog $\dfrac{\log \frac{P_1}{P_0} \; 100}{N}$

where N refers to the number of items whose price relatives are thus averaged.

Illustration No. 2 Calculate Index Numbers for 1977, 1979 and 1981 taking 1975 as base from the following data by Mean of Relative Method.

Commodity	1975	1977	1979	1981
A	2	5	4	3
B	8	11	13	6
C	4	5	6	8
D	6	4	5	7
E	5	4	6	3

Solution :

Construction of Index Numbers based on Mean of Relatives.

Commodity	1975		1977		1979		1981	
		P_0	P_1	$\frac{P_1}{P_0}$ 100	P_2	$\frac{P_2}{P_0}$ 100	P_3	$\frac{P_3}{P_0}$ 100
A	2	100	5	250.0	4	200.0	3	150.0
B	8	100	11	137.5	13	162.5	6	75.0
C	4	100	5	125.0	6	150.0	8	200.0
D	6	100	4	66.7	5	83.3	7	116.7
E	5	100	4	80.0	6	120.0	3	60.0
		500		659.2		715.8		601.7
				$\frac{P_1}{P_0}$ 100		$\frac{P_2}{P_0}$ 100		$\frac{P_3}{P_0}$ 100

(with usual symbols and A.M. as average)

P_{01} = Index with 1975 as base and 1977 as current years $\frac{359.2}{5}$ 13183

P_{02} = Index with 1975 as base and 1979 as current years $\frac{715.8}{5}$ 143.16

P_{03} = Price Index with 1975 as base and 1981 as current years $\frac{601.7}{5}$ 120.33.

2. Weight Index Number

%4 u Aggregative Method: These indices are of the simple aggregative type with the only differences that the weights are assigned to the various items included in the index. Since there are various methods of assigning weights, there are corresponding various formulas for the calculation of Index Numbers.

This method in fact can be described as an extension of the simple aggregative method in the sense that the weights are assigned to the different commodities in the index. There are various methods by which weights can be assigned and hence a large number of formulae for constructing Index Numbers have been devised. Some of the various methods suggested by different authorities are as follows:

%4 u Laspeyre's method.

%4 u Paasche's method.

%4 u Fisher's ideal method.

%4 u Marshall Edge worth method.

%4 u Kelly's method.

%4 u Dorbish and Bowley's method.

%4 u Laspeyre's Method

Lespeyre suggested that for the purpose of calculating Price Indices, the quantities in the base year should be used as weights. Hence the formula for price Index Number according to this methods would be:

$$P_{01} \quad \frac{p_1q_0}{p_0\,q_0} \; 100$$

Where	P	refers to Price Index
	p	refers to Price of each commodity.
	q	refers to Quantity of each commodity
	0	base year,
	1	current year, and
		refers to the summation of the items.

In other words the step for calculating Index Numbers according to this method are:

%4 u Multiply the price of each commodity for current year with its respective Quantity of that Commodity for the base year ($p_1\,q_0$) and then find out the total of this product (p_1q_0).

%4 u Multiply the price of each commodity for the base year with the respective quantity of the commodity for the base year ($p_0\,q_0$) and then find out the total of these products for different commodities ($p_0\,q_0$).

%4 u Divide ($p_1 . q_0$) with ($p_0 . q_0$) and multiply the quotient by 100. These gives us the Price Index. On the other hand, if Quantity Index by this method is to be calculated, the prices in the base year will be used weights.

Symbolically

$$Q_{01} \quad \frac{q_1\,p_0}{P_0\,q_0} \; 100$$

(symbols have the same meaning as stated in the earlier portion).

Illustration 3. Compute Price Index and Quantity Index from data given below by Lespeyre's method.

Items	Base year		Current year	
	Quantity	Price	Quantity	Price
A	6 units	40 paise	7 units	30 paise
B	4 units	45 paise	5 units	50 paise
C	5 units	90 paise	1.5 units	40 paise

Solution : Computation of Price and Quantity Indices.

Items	Base year		Current year					
	q_0	p_0	q_1	P_1	p_0q_0	p_1q_1	p_0q_1	p_1q_1
A	6	40	7	30	240	180	280	210
B	4	45	5	50	180	200	225	250
C	5	90	1.5	40	450	200	135	60
				Total	p_0q_0 = 870	p_1q_0 = 580	p_0q_1 = 640	p_1q_1 = 520

$$P_{01} \frac{p_1q_0}{p_0q_0} 100 \qquad \frac{580}{870} 100 \;\; 66.66.$$

and Quantity index or $Q_{01} \frac{q_1p_0}{q_0p_0} \; \frac{640}{870} 100 \qquad \frac{640}{465} 100 \quad 73.56$

%4 u Paasches Method. Under this method of calculating Price Index the quantities of the current year are used as weights as compared to base year quantities as suggested by Lespeyre.

Symbolically Price Index or $P_{01} \frac{p_1q_1}{p_0q_1} 100$

Hence the steps of construction according to Paasche's method are:

%4 u Calculate the product of the current year prices of different commodities and their respective quantities for the current year (p_0 q_1) and find out the total of the product of different commodities (p_1 q_1)

%4 u Calculate the product of p_0 and q_1 of different commodities and total them up. (p_0q_1) .

%4 u Divide on (p_1 q_1) with (p_0q_1) and multiply the quotient by 100 to obtain price Index.

In the same manner if quantity index by this method is to be calculated the current year price are used as weights. Symbolically $Q_{01} \frac{p_1 \;\; q_1}{p_0 q_1} 100$

Illustration 4. For the data given in illustration 3 calculate (i) Price Index (ii) Quantity Index by using Paasche's method.

Items	Base year		Current year					
	q_0	p_0	q_1	P_1	p_1q_0	p_0q_1	p_1q_1	p_1q_1
A	6	40	7	30	240	180	280	210
B	4	45	5	50	180	200	225	250
C	5	90	1.5	40	450	200	135	60
				Total	p_0q_0 = 870	p_1q_0 = 580	p_0q_1 = 640	p_1q_1 = 520

Price Index $P_{01} \frac{p_1 q_1}{p_0 q_1} 100 \qquad \frac{520}{640} 100 \;\; 81.5.$

Quantity Index or $Q_{01} \frac{q_1 p_1}{q_0 p_1} \;\; 100 \;\; \frac{520}{580} 100 \quad 89.65$

%4 u Fisher's Ideal Index: Lespeyre has used Base year quantities as weight where as Paasche's has used current year quantities as weights for the computation of Index Number of prices. Fisher suggested that it is both the current year quantities as also the year quantities that should be used but geometric mean is to be calculated and that figure should be the Index Number. Symbolically,

Fisher's price Index

$$= P_{01} \quad \sqrt{\frac{p_1 q_0}{p_0 q_0} 100 \quad \frac{p_1 q_1}{p_0 q_1} 100}$$

$$\sqrt{\frac{p_1 q_0}{p_0 q_0} \quad \frac{p_1 q_1}{p_0 q_1}} \quad 100$$

Fisher's method is also expressed as Fisher's Index

$$\sqrt{\text{Lespeyre's Index} \text{Paasches's Index}}$$

On the other hand if quantity Indices by this method are to be calculated the geometric mean of the Index Number of quantities with base year prices as weights and Index Number of Quantities with current year as weights be found out. Symbolically,

Fisher's Quantity Index = Q_0 $\sqrt{\frac{q_1 p_0 \quad q_1 p_1}{q_0 p_0 q_0 p_1}} \ 100$

Illustration 5. Construct Index Number of Prices and Quantities from the following data using Fisher's method 11980=100).

Commodity	1980		1982	
	Price	Qty.	Price	Qty.
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution : calculation of Price and Production Indices.

	1980		1982					
Commodities	Price	Qty.	Price	Qty	p_0q_0	p_1q_1	p_1q_0	p_0q_1
A	2	8	4	6	16	24	32	12
B	5	10	6	5	50	30	60	25
C	4	14	5	10	56	50	70	40
D	2	19	2	13	38	26	38	26
					160	130	200	103

$$P_{01} \quad \sqrt{\frac{p_1 q_0}{p_0 q_0} \quad \frac{p_1 q_1}{p_0 q_1}} \ 100 \qquad \sqrt{\frac{200}{160} \quad \frac{130}{103}} \qquad 100 \quad 125.6.$$

$$Q_{01} \quad \sqrt{\frac{q_1 p_0}{q_0 p_0} \quad \frac{q_1 p_1}{q_0 p_1}} \ 100 \qquad \sqrt{\frac{103}{160} \quad \frac{130}{200}} \ 100 \quad 64.7.$$

%4 u Marshall & Edge worth’s Method. In this method also current year as well as base year prices and quantities are considered. The formula (with usual notations) is as follows:

$$P_{01}\ \frac{(q_0+q_1)\,p_1}{(q_0+q_1)\,p_0}\,100 \qquad \frac{q_0\,p_1 q_1\,p_1}{q_0\,p_0 q_1 p_0}\,100$$

and Quantity Index is calculated by the formula

$$Q_{01}\ \frac{(p_0\ +p_1)\,q_1}{(p_0\ +p_1)\,q_0}\,100 \qquad \frac{p_0\,q_1 p_1\,q_1}{p_0\,q_0 p_1 q_0}\,100$$

(5) Kelly’s Method. Truman Kelly has suggested the following formula for constructing Index Number.

$$P_{01}\,\frac{p_1\,q}{p_0\,q}\,100$$

$$\text{where } q\ \frac{q_0+q_1}{2}$$

where ‘q’ refers to the average quantity of some period. This method is also known as the fixed aggregative method.

%4 u Dorbish & Bowley’s Method. Dorbish & Bowley have suggested the simple arithmetic mean of lespeyre’s and Paasche’s formula. Symbolically,

$$P01\,\frac{\frac{p_1\,q_0}{p_0\,q_0}\quad\frac{p_1\,q_1}{p_0\,q_1}}{2}\ 100.$$

Self Assessment Multiple Choice Questions:

%4 u The Laspeyres's and Paasche's formulae are generally preferred for the construction of

(a) Index numbers (b) Frequency table (c) Pie charts (d) Bar graphs

7. Index numbers are expressed in terms of

(a) Constant (b) Decimals (c) Percentages (d) Decimals

8. Laspeyres's index requires .... Calculation work than the one with changing weights in every period.

(a) Simple(b) Complex (c) Less (d) More:

%4 u In practical situations, neither .... nor .... change in the same proportion, the two index numbers are in general different from each other.

(a) Prices, quantities (b) Size, Volume (c) Price, Value(d) Ratio, Quantity

%4 u Weighted Average of Price Relatives:

This method is also known as the ‘Family Budget Method’. Weights are values of the base year in this method. The Index Number for the current year is calculated by dividing the sum of the products of the current year’s price relatives and base year values by the total of the weights, i.e., the weighted Arithmetic average of the price relative gives the required index numbers. Symbolically.

Weighted Index Number of the Current year

%4 u $\frac{IV}{V}$

where I stands for Price Relatives of the current year, and ‘V stands for the values of the base year.

Illustration 6: From the data given below, calculate the Weighted Index Number by using Weighted Average of Relatives.

Commodities	Units	Base Yr. Qty.	Base Year's Price	Current Year's Price
A	Quintal	7	16	19.6
B	Kg.	6	2	3.2
C	Dozen	16	5.6	7.0
D	Meter	21	1.5	1.4

Solution :

The Price relative of the current year $\frac{\text{Current Year's Price}}{\text{Base Year's Price}}$ **100**

The value of the Base year = Quantity of Base year Price of the Base year

commodities	Price Relatives of the Current Year i.e. I $\frac{p_1}{p_0}$ 100	Values or Weights i.e. V = p_0q_0	Weights Price Relatives V I
A	122.5	112.0	13,720
B	160.0	12.0	1,920
C	15.0	89.6	11,200
D	93.3	31.5	2,939
		V = 245.1	IV = 29,779

Weighted Index Number of Prices = $\frac{IV}{V}$

Weighted Index Number of Prices $\frac{29779}{245}$

Weighted Index Number of Prices = 121.5 Ans.

In weighted average of relatives, the Geometric mean may be used instead of arithmetic mean. The weighted geometrical mean of relatives is calculated by applying logarithms to the relatives. When this mean is used, then the formula is:

$$P_{01} = \text{Antilog} \frac{V.\log I}{V}$$

where I $\frac{P_1}{p_0}$ 100 , V = p_0q_0 or value weights.

Illustration 7 : Find out price index by weighted average of price relative from the following commodities us geometric mean :

Commodities	p_0	q_0	p_1
X	3.0	20	4.0
Y	1.5	40	1.6
Z	1.0	10	1.5

Solution :

Calculation of Index Number

Commodities	p_0	q_0	p_1	V (p_0q_0)	$\frac{P_1}{p_0}$ 100 I	Log I	V. log I
X	3.0	20	4.0	60	$\frac{4}{3}$ 100 133.33	2.1249	126.494
Y	1.5	40	1.6	60	$\frac{1.6}{1.5}$ 100 106.7	2.0282	121.692
Z	1.0	10	1.5	10	$\frac{1.5}{1.0}$ 100 150.0	2.1761	21.761
			V	= 130		V. log 1 = 270.947	

By applying the formula :

$$P_{01} \quad AL\frac{V.\log I}{V} \quad AL\frac{270.947}{130} \quad AL\ 2.084 \text{ i.e. } =121.3.$$

Tests of Adequacy of Index Numbers

Since several formulas have been suggested for the construction of index numbers, then the question arises which method of index number is the most suitable in a given situation. These are some of tests to choose an appropriate index:

%4 u Unit Tests : It requires that the method of constructing index should be impendent of the units of the problem. All the method except simple aggregative method satisfy this test.

%4 u Circular Test : This test was suggested by Westerguard and C.M. Walsch. It is based on the shiflability of the base. Accordingly the index should work in a circular fashion i.e., in an index number is computed for the period. On the base period 0, another index is computed for 2 on the base period 1, and still another index number is computed for period 8 on the base period 2. Then the product should be equal to one.

i e P_{01} P_{12} P_{23} .... P_{n0} = 1.

Only simple aggregative and fixed weight aggregative method satisfy the test. If the test is applied to simple aggregative method, we will get.

$$\frac{p_1}{p_2}\ \underline{p_2\ p_3}\ 1\ \underline{p_0\ p_1}$$

The test is met by simple geometric mean of price relatives and the weighted aggregative fixed weights.

%4 u Time Reversal Test : According to Prof. Fisher the formula for calculating an index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as the base. In other words, when the data for any two years are treated by the same method, but with the base reversed, the two index numbers should be reciprocals of each other. In symbols

P_{01} P_{10} = 1.

(omitting the factor 100 from each other)

Where P_{01} denoted the index for current year 1 based on the base year 0 and P_{10} is for current year 0 on the base year 1.

It can be easily verified that simple geometric mean of price relative index. Weighted aggregative formula, weighted geometric mean of relative and Marshall Edge worth and Fisher's ideal methods satisfies this test.

Let us see how Fisher ideal method satisfied the test.

$$P_{01} \sqrt{\frac{p_1 q_0}{p_0 q_0} \quad \frac{p_1 q_1}{p_0 q_1}}$$

By changing time from 0 to 1 to 0

$$P_{10} \sqrt{\frac{p_0 q_1}{p_1 q_1} \quad \frac{p_0 q_0}{p_1 q_0}}$$

Now $P_{01} \quad P_{10} = 1.$

Substitute the value of P_{01} and P_{10}

$$P_{01} \; P_{10} \quad \sqrt{\frac{p_1 q_0}{p_0 q_0} \quad \frac{p_1 q_1}{p_0 q_1}} \quad \frac{p_0 q_1}{p_1 q_1} \quad \frac{p_0 q_0}{p_1 q_0} \; 1.$$

%4 u Factor Reversal Test: It was that the product of a price index and the quantity index should be equal to value index. In the words of Fisher, just as each formula should permit the interchange of the two times without giving inconsistent results so it should permit interchanging the prices and quantities without giving inconsistent result which means the two results multiplied together should give the true value ratio. The test says that the change in price multiplied by change in quantity should be equal to total change in value. Analytically if P_{01} is a price index for the current year reference to base year.

Q_{01} is the quantity index for the current year.

Then $Q_{01} \quad P_{01} \quad \dfrac{p_1 q_1}{p_0 q_0}$

This test is satisfied only by Fisher's ideal index method.

$$P_{01} \sqrt{\frac{p_1 q_0}{p_0 q_0} \quad \frac{p_1 q_1}{p_0 q_1}}$$

Changing p to q and q to p.

$$Q_{01} \sqrt{\frac{q_1 p_0}{q_0 p_0} \quad \frac{q_1 p_1}{q_0 p_1}}$$

$$P_{01} \; Q_{01} \quad \sqrt{\frac{p_1 q_0}{p_0 q_0} \quad \frac{p_1 q_1}{p_0 q_1} \quad \frac{q_1 p_0}{q_0 p_0} \quad \frac{q_1 p_1}{q_0 p_1}} \quad \sqrt{\frac{(p_1 q_1)^2}{(p_0 q_0)^2}} \quad \frac{p_1 q_1}{p_0 q_0}$$

In other words, factor reversal test is based on the following analogy. If the price permit of commodity increases from Rs. 10 in 1975 to Rs. 15 in 1978, and the quantity of consumption change from 100 units to 140 units during the same period, the then price and quantity in 1978 are 150 and 140 respectively. The values of consumption (p q) were Rs. 1000 in 1975 and Rs. 2100 in 1978 giving a value ratio. $\dfrac{2100}{1000}$ 2.1

Thus we find that the product of price ratio and quantity ratio equals the value ratio :

1.5 x 1.4 = 2.1

Chain Base Index

The various formulas discussed so far assume that base period is some fixed previous period. The index of a given year on a given fixed base is not affected by changes in the prices or the quantities of any other year. On the other hand, in the chain base method, the value of each period is related with that of the immediately proceeding period and not with any fixed period. To construct index numbers by chain base method, a series of index numbers are computed

for each year with proceeding year as the base. These index numbers are known as Link relatives. The link relatives when multiplied successively known as the chaining process give relative to a common base. The products obtained are expressed as % and give the required index number by this method. The required steps of chain index are:

(i) Express the figures of each period as a % of the proceeding period to obtain Link Relatives (LR).

(ii) These link relatives are chained together by successive multiplication to get chain indices by the formula:

Chain Base index (CBI)

$$= \frac{\text{Current year LR Preceding year's Chain Index}}{100}$$

%4 u The chain index can be converted into a fixed base index by this formula: Current year Fixed Base Index (FBI)

$$\frac{\text{Current year's CBI Previous year's FBI}}{100}$$

Chain relatives are computed from link relatives whereas fixed base relatives are computed directly from the original data. The results obtained by fixed base and chain base index invariably are the same.

We shall illustrate the process by taking some examples.

Illustration 8: Construct Index Numbers by chain base method from the following data of wholesale prices of cotton.

Year :	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Price:	75	50	65	60	72	70	69	75	84	80

Solution:

Calculation for Chain Index

Year	Price	Link Relatives	Chain Base Index	Fixed Base Index
1971	75	100	100	100
1972	50	$\frac{50}{75}$ 100 66.67	$\frac{66.67\ \ 100}{100}$ 66.67	$\frac{50}{75}$ 100 66.67
1973	65	$\frac{65}{50}$ 100 130.00	$\frac{130\ \ 66.67}{100}$ 86.67	$\frac{65}{75}$ 100 86.67
1974	60	$\frac{60}{65}$ 100 92.31	$\frac{92.31\ \ 86.67}{100}$ 80.00	$\frac{60}{75}$ 100 80.00
1975	72	$\frac{72}{60}$ 100 120.00	$\frac{120\ \ 80}{100}$ 96.00	$\frac{72}{75}$ 100 96.00
1976	70	$\frac{70}{72}$ 100 97.22	$\frac{97.22\ \ 96}{100}$ 93.33	$\frac{70}{75}$ 100 93.33
1977	69	$\frac{69}{70}$ 100 98.57	$\frac{98.57\ \ 93.33}{100}$ 92.00	$\frac{69}{75}$ 100 92.00
1978	75	$\frac{75}{69}$ 100 108.69	$\frac{108.69\ \ 92}{100}$ 100.00	$\frac{75}{75}$ 100 100.00
1979	84	$\frac{84}{75}$ 100 112.00	$\frac{112\ \ 100}{100}$ 112.00	$\frac{84}{75}$ 100 112.00
1980	80	$\frac{80}{84}$ 100 95.24	$\frac{95.24\ \ 112}{100}$ 106.67	$\frac{80}{75}$ 100 106.67

Solution :

Calculations for Chain Index

Year	Link Relatives	Chain Index Number
1975	100	100
1976	105	$\frac{105}{100}$ 100 105.00
1977	95	$\frac{95}{100}$ 105 99.75
1978	115	$\frac{115}{100}$ 99.75 114.7
1979	102	$\frac{102}{100}$ 114.75 137.04

Base Shifting: Sometimes it becomes necessary to change the base of index series from one period to another for the purpose of comparison. In such circumstances it is necessary to recomputed all index numbers using new base period. Such computation of index numbers using new base period is to divide index number in each period by the index number corresponding to the new period and then to express the result as percentage. This process is known as chaining the base.

Illustration 10. Compute Index Numbers from the following taking 1976 as the base and shift the base to 1978.

Year:	**1976**	**1977**	**1978**	**1979**	**1980**
Price:	**10**	**12**	**15**	**21**	**20**

Solutoon :

Calculations for Chain Index

Year	Price	Index Number Base 1976	Shift to base from 1976 to 1978
1976	10	100	$\frac{100}{150}$ 100 67
1977	12	$\frac{12}{10}$ 100 120	$\frac{120}{150}$ 100 80
1978	15	$\frac{15}{10}$ 100 150	100
1979	21	$\frac{21}{10}$ 100 210	$\frac{210}{150}$ 100 140
1980	20	$\frac{20}{10}$ 100 200	$\frac{200}{150}$ 100 133

It may be seen that index by chain base and fixed base method comes to the same.

Illustration 9. Construct chain index numbers from the link relative given below:

Year:	1975	1976	1977	1978	1979
Link Relatives:	100	105	95	115	102

Splicing: On several occasions the base year may give discontinuity in the construction of index numbers. We would always like to compare figures with a recent year and not with a distant past. For example, the weights of an Index number may become out of data and we may construct another index with new weights. Two indices would appear. It becomes necessary to convert these two indices into a continuous series. The procedure employed to do the conversion is known as splicing. The formulae are :

For Forward Splicing :

Spliced Index Number: $$\frac{\text{Old index of the New Base year Index to be Adjusted}}{100}$$

Spliced Index Number : $$\frac{100}{\text{Old Index of New Base year}}\ \text{Index to be Adjusted}$$

Illustration 11. Splice the following two Index Number series continuing series A forward and the series B backward :

Year:	1975	1976	1977	1978	1979	1980
Series A :	100	120	150	—	—	—
Series B:	—	—	100	110	120	150

Solution :

Splicing of two Index Number Series

Year	Series A	Series B	Index Numbers Spliced forward to Series B	Index Numbers Spliced forward to Series A
1975	100		$\frac{100}{150}$ 100 66.66	
1976	120		$\frac{100}{150}$ 120 80.00	
1977	150	100	$\frac{100}{150}$ 150 100.00	$\frac{150}{100}$ 100 150
1978		110	$\frac{100}{150}$ 110 165	
1979		120	$\frac{100}{150}$ 120 180	
1980		150	$\frac{100}{150}$ 150 225	

Deflating : It means making allowance for the change in the purchasing power of money due to a change in general price level. It is the technique of converting a series of values calculated at current prices into a series at constant prices of a given year. In other words the process of removing the effects of price changes from the current money values is called Deflation. By this process that real value of the phenomenon is calculated which is free from the influence of price changes. Deflation is used in computation of national income and other economic variables. The relevant price index is called the deflator whether it is to be the wholesaler price index or consumer price index. Normally separate price deflators are found one for deflating the national income data from different sectors to the economy considering the changes in prices in those sectors because the general sectors of the economy considering the changes in prices in those sectors. Because the general price rise in agriculture products may be more than in the individuals products. The method is:

$$\text{Deflated value} = \frac{\text{Current Price (value)}}{\text{Deflator}} \quad 100$$

Consumer Price Index Numbers

The consumer price index known as cost of living index is calculated to know the average changes over time in the prices of commodities consumed by the consumers. The need to construct consumer price indices arises because the general index numbers fail to give an exact idea of the effect of the change in the general price level on the cost of living of different classes of people, because a given changes in the level of prices affect different classes of people in different manners. Different people consume different kinds of commodities and if same commodities, in different proportions. The consumer price index helps us in determining the effect of rise and fall in prices on different classes of consumers living in different areas. The consumer price index is significant because the demand of a higher wage is based on the cost of living index and the wages and salaries in most nations are adjusted according to this index. We should understand in the cost of living index does not measure the actual cost of living nor the fluctuations in the cost of living due to causes other than the change in price level but in object is to find out how much the consumers of a particular class have to pay more for a certain basket of goods and services. That is why the term cost of living index has been replaced by the term price of living index, cost of living price index or consumer price index.

The significance of studying the consumer price index is that it helps it wage negotiations and wage contracts. It also helps in preparing wage policy, price policy, rent control, taxation and general economic policies. This index is also used to find out the changing purchasing power of the currency.

The consumer price index can be prepared by two methods:

%4 u Aggregative Method;

%4 u Weighted Relative Method.

When aggregative method is used to prepare consumer price index, the aggregative expenditure for current year and base year are calculated and the below given formula is applied.

$$\text{Consumer Price Index} = \frac{p_1 q_0}{p_0 q_0} \quad 100$$

When weighted relatives method is used then the family budgets of a large number of people for whom the index is meant are carefully studied and the aggregative expenditure of an average family on various items is estimated. These will be weights. In other words, the weights are calculated by multiplying the base year quantities and prices (p_0q_0)

The price relatives for all the commodities are prepared and multiplied by the weights. By applying the formula, we can calculate Consumer Price Index.

C.P.I. $\underline{IV}$

V

$$I \quad \frac{p_1}{p_0} \; 100 \quad V = p_0 q_0 .$$

Illustration 12. Prepare the consumer price index for 1981 on the basis of 1980 from the following data by both methods.

Commodities	Quantities Consumed in 1980	Price in 1980	Price in 1981
A	6	5.75	6.00
B	6	5.00	8.00
C	1	6.00	9.00
D	6	8.00	10.00
E	4	2.00	1.50
F	1	20.00	15.00

Solution :

(i) Consumer Price Index by Aggregative Method

Commodities	q_0	p_0	P_1	P_1q_0	p_0q_0
A	6	5.75	6.00	36.00	34.50
B	6	5.00	8.00	48.00	30.00
C	1	6.00	9.00	9.00	6.00
D	6	8.00	10.00	60.00	48.00
E	4	2.00	1.50	6.00	8.00
F	1	20.00	15.00	15.00	20.00
				P_1q_0 =174	p_0q_0 = 46.5

Consumer Price Index $\frac{p_1q_0}{p_0q_0}$ 100 $\frac{174}{146.5}$ 100 $118.77.$

(ii) Consumer Price Index by Weighted Relatives

Commodities	q_0	p_0	p_1	I	V	IV
A	6	5.75	6.00	104.34	34.5	3600
B	6	5.00	8.00	160.00	30.0	4800
C	1	6.00	9.00	150.00	6.0	900
D	6	8.00	10.00	125.00	48.0	6000
E	4	2.00	1.50	75.00	8.0	600
F	1	20.00	15.00	75.00	20.0	1500
				V = 146.5		V= 17400

Consumer Price Index = $\frac{IV}{V}$ = $\frac{17400}{146.5}$ $118.77.$

Index Number of Industrial Production

The Index Number of industrial production is prepared to know the increase or decrease in the level of industrial production in a given period compared with some other period. This index measures the changes in quantum of production. To prepare such an index it is necessary for us to compute the production for two periods i.e., for the current year and for the base year. Generally the data are collected under these heads :

%4 u Textile industries—cotton, woolen, silk etc.

%4 u Mining industries—iron-ore, coal, copper, petroleum etc.

%4 u Metallurgical industries—iron and steel,

%4 u Mechanical industries—locomotives, ships, aeroplanes etc.

%4 u Miscellaneous—glass, soap, chemical, cement etc.

The output for various industries are computed. Weights are assigned to various industries on the basis of some criteria as capital invested, turnover, net output, production etc. We apply this formula:

Index of Industrial Production

$\frac{IW}{W}$ where I $\frac{q_1}{q_0}$

W = Relative importance of different outputs.

Limitation of Index Numbers

%4 u They are only approximate indicators of the relative level of a phenomenon.

%4 u Index numbers are good for one purpose may be unsuitable for the-other.

%4 u Index numbers can be manipulated in such a manner as to draw the desired conclusions.

(C) Self Assessment Fill in the blanks:

10. While taking weighted average of price relatives, the .... are often taken as weights.

11. Laspeyres's Index has an .... bias.

12. Paasche's Index has a .... bias.

13. Weighted aggregative and weighted arithmetic average of price relatives, are ....

14. One type of index number can be obtained from the other by .... of weights.

15. The weighted aggregative index numbers are .... to calculate and have .... Interpretation.

State whether the following statements are true or false :

%4 u Index numbers where comparisons of various periods were done with reference to a particular period, termed as base period.

%4 u There is no problem with a fixed base series even when the base year becomes too distant from the current year.

%4 u When there is a single commodity, the chained index will be equal to the fixed base index.

Self Check Exercise :

%4 u What are index numbers?

%4 u What are time reversal and factor reversal tests?

6.8. Summary

Index Numbers are indicators which reflect the relative changes in the level of a certain phenomenon in any given period (or over a specified period of time) called the current period with respect to its values in some fixed period, called the base period selected for comparison. Though originally designed to study the general level of prices, today index numbers are extensively used for a variety of purposes in economics, business, management etc, and for quantitative data relating to production, consumption, profits, personnel and financial matters, etc., for comparing changes in the level of phenomenon for two periods, places, etc. Index numbers are the Economics barometers, they help in studying trends and tendencies, formulating decisions and policies.

6.9 Glossary:

Index number : An index number is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations.

Base Year: The year from which comparisons are made is called the base year. It is commonly denoted by writing '0' as a subscript of the variable.

Current Year: The year under consideration for which the comparisons are to be computed is called the current year. It is commonly denoted by writing ' 1' as a subscript of the variable.

Barometers of economic activity: Sometimes index numbers are termed as barometers of economic activity.

Dorbish and Bowley's index: This index number is constructed by taking the arithmetic mean of the Laspeyres's and Paasche's indices.

Fisher's Index: Fisher suggested that an ideal index should be the geometric mean of Laspeyres' and Paasche's indices.

Laspeyres's Index: Laspeyres' price index number uses base year quantities as weights

Paasche's Index: This index number uses current year quantities as weights.

Quantity Index Number: A quantity index number measures the change in quantities in current year as compared with a base year.

Simple Aggregative Method: In this method, the simple arithmetic mean of the prices of all the items of the group for the current as well as for the base year are computed separately. The ratio of current year average to base year average multiplied by 100 gives the required index number.

Value index Number: A value index number gives the change in value in current period as compared with base period. The value index is denoted by V01. for all periods Weighted Aggregative Method: This index number is defined as the ratio of the weighted arithmetic means of current to base year prices multiplied by 100.

6.10 Answers: Self Assessment

1. Index Number **2. characteristics**

3. extent, direction **4. base**

5. current **6. (a)**

7. (C) **8. (C)**

9. (a) **10. values**

I1. upward **12. downward**

13. same **14. Suitable selection**

15. easy, simple **16. True**

17. False **18. True**

19. Refer to section 6.3 **20. Refer to section 6.6**

6.11 Terminal Questions

6.11. Terminal Questions:

%4 u What do you mean by Index Numbers? Discuss the various problems related to Index Numbers.

2. What is Fisher's Ideal Index? Why is it called Ideal?

%4 u Discuss the properties of Laspeyre's and
fisher's Index Number. 6.12 Suggested Readings

Hooda, R.P., Statistical Methods, MacMillan, New Delhi.

Gupta, S.P., Statistical Methods, Sultan Chand & Sons,

Gupta, S.C., Business Statistics, Himalaya Publishing House, Mumbai.

*****

Lesson-7
Probability

Structure

7.1. Learning Objectives

After Studying this Lesson, you should be able to understand:-

%4 u Probability Theory
%4 u Different types of Events
%4 u Different Approaches to the theory of Probability
%4 u Probability Rules
%4 u Bayes' Theorem

7.2. Introduction

Many phenomenons of business, economics and social sciences possess uncertainty on the outcomes may not always be the same even under the same set of circumstances. They lack deterministic regularity in its outcomes. For instance, the number of accidents on a highway in the month of January may be quite different from the number of accidents in the month of February though the circumstances were similar. Uncertainty is a part and parcel of decision making in social phenomenons. Weather forecasting, share prices forecasting, product quality, number of defective parts produced by a machine, monsoons in the next year, customers reaction towards a new product, voter's preference for a particular candidate etc. are some of the areas, where, commenting on the future outcomes with certainty becomes impossible. All these examples illustrate that future is uncertain and uncertainty is an established fact of life. In these situation, probability theory helps us with means and ways to attain the precise expressions for uncertainties involves.

In our day to day life, we also use the concept of probability, for example, it might rain, today, probable the price of X share will go up, probably 'X' team will win the match, there are good chances that our sales will go up by 10 per cent next year, our company may cross the target profit this year. All these terms may, probably, possibly, likely etc. convey the same sense i.e. the events are not certain to take place. It connotes the sense that there is uncertainty about the happening of an event.

The theory of probability has its origin in the game of chances relating to gambling such as tossing of a coin throwing a dice, drawing a card from a pack of cards etc. Over the years, different authorities contributed in developing the probability theory namely Galileo, Pascal, Fermat, Bernoulli, Abraham De Moivre. Thomas Bayer, Laplace, Fisher, Von Mises, Markov, Kolmogorov etc. From the games of chances, probability today has become the basic tool in statistics. Probability theory is being applied in decision making in economics, business, political, social, medicine, insurance etc.

It order to make statements about the uncertain environment, we need to develop a language. Probability is a language in which we discuss uncertainty. Before one can communicate with another in this language, one should have a common vocabulary. It is necessary to introduce a terminology used in probability.

7.3. Experiment, Outcomes, Events and Sample Space:

Probability in common Parlance, connotes the chances of occurrence of an event of happening. A process which could lead to two or more than two different outcomes and there is uncertainty as to which outcome will occur is called experiment. For example:

%4 u tossing of a coin

%4 u rolling a dice

%4 u voters preference for a candidate

%4 u launching of a new product in the market

%4 u asking worker whether he has joined the trade union. All these processes have three features:

%4 u There are two or more than two possible outcomes;

%4 u It is possible to specify all the outcomes in advance.

%4 u There is uncertainty about the outcomes or happening.

Any process which have all the three features is called a random experiment in probability theory terminology. For example tossing of a coin results in either head or tail. .Both these outcomes can be determined in advance and there is uncertainty which outcome will occur. In the case of dice, the outcomes will be either 1, 2, 3, 4, 5, or 6. A voter might show a preference for Candidate X or Y or Z. In each, the different possible outcomes are called basic outcomes, have been listed. The collection of possible outcomes of an experiment is called sample space or the set of all possible outcomes is called the sample space of an random experiment. In a single coin the sample space has two points.

S=[H,T]

In case of a die, the sample space has six points.

S=[1,2,3,4,5,6]

In questioning the worker about union membership, the sample space has two points—Yes or No.

Thus, the sample space is determined by listing down all the possible outcomes of an random experiment. Frequently, one is interested not in the basic outcomes themselves but in some subset of all the possible in sample space. For example, if a, die is rolled, the odd number is defined as success and even a failure. The success can occur if number 1, 3 or 5 occurs an failure if number 2, 4 or 6 turns up. Such set of basic outcome are called events.

Consider the experiment of testing three successive unit from the production lot. The possible outcomes can be

Even	Unit I	Unit II	Unit III	Outcome
0 Defective	G	G	G	GGG
1 defective	D	G	G	DGG
	G	D	G	GDG
	G	G	D	GGD
2 defective	G	D	D	GDD
	D	G	D	DGD
	D	D	G	DDG
3 defective	D	D	D	ODD

D is defective and G is good.

The above table depicts all possible eight outcomes of experiment. The sample space can be written as:

S = [GGG, GDG, GGD, GDD, DGD, DDG, DDD]

In another example, let X is the sum of the spots on the s of two ideal dice. The set of the all possible outcomes of experiment will be as follows:

Second dice

		1	2	3	4	5	6
	1	2	3	4	5	6	7
dice	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
First	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

The six faces of the one dice is represented on one side and those of the other on the top row. There are 6 6 = 36 possible outcomes is this experiment. These 36 outcomes are elementary outcomes and they can be grouped into events. The corresponding values of which form the sample space

s = [2, 3, 4, 5, 6, 7, 8, 9, 10,11,12]

Event 2 (total) occurs only one way i.e. 1 in each dice, 3 occurs two ways, 4 occurs three ways and so on. Thus event is a subset of the sample space. Unlike the common usage of the term, where an event refers to a particular happening or incident. But in probability theory, event refers to a single outcome or a combination of outcomes. One must know the number of elementary outcomes that make up the particular event. In the roll of two dice the total seven (event) can occur in six ways outcomes, namely (1,6), (2,5), (3,4), (4,3), (5,2) and (6,1)

7.4. Approaches to Probability Theory:

Suppose a random experiment is carried out and we are interested in knowing the occurrence of a particular event. The concept of probability provides a numerical value for the likelihood of an events occurrence. Probability is measured on a scale from 0 to 1. Probability of 0 implies that event is impossible or it is certain it will not occur. On the other extreme, probability 1 implies, it is certain that event will happen. For uncertain events, we assign the probability between 0 and 1. Higher the probability, more likely the event to occur. Thus, probabilities are expressed as fractions (1/2,1/3,1/4, 3/4,5/9) or in decimals (0.215, 0.345).

There are three basic approaches to study the probability theory. These are;

%4 u Classical approach

%4 u Relative frequency approach

%4 u Subjective approach

7.4. (1) Classical Approach :

The classical approach to probability is the oldest and simplest one. It originated from the game of chances like tossing of a coin, throwing a dice, pack of cards etc. It is based on considerations of symmetry and logic.

The basic assumption in classical approach is, the outcomes are 'equally likely' e.g. tossing a coin there are equal chances of Head and Tail. In case of a dice, there are equal chance of six outcomes i.e. 1, 2, 3, 4, 5, and 6. Similarly in a pack of cards, there are 52 outcomes and each outcome has equal chances of coming up. There is no reason for believing one outcomes to be more likely than another.

Classical approach defines the probability that an event will occur as follows

$$\text{Probability of occurrence of an event} = \frac{\text{Number of favourable outcomes}}{\text{Total number of equally likely outcomes}}$$

For calculating probability, one must have the information

%4 u number of favourable outcomes where event can occur.

%4 u total number of equally likely outcome.

Suppose a coin is tossed, the probability of head in a single will be P (head) $\frac{1}{1\;1}\quad\frac{1}{2}$

Head outcome can occur in only one way and total number of outcomes are two. Similarly in dice roll experiment, the probability of getting 3 is

$$P(3)\quad \frac{1}{1\,1\;\;1\;\;1\;1\;1}\quad \frac{1}{6}$$

From a pack of cards, one card is drawn. Probability of getting a black card from a pack is $\frac{26}{52}$. Probability of getting a spade is $\frac{13}{52}$. Portability of getting a king is $\frac{4}{52}$, Probability of getting a Queen of hearts is $\frac{1}{52}$.

Classical approach is also called a priori probability as one can state the answer in advance (a priori) without doing experiment. The answer is based on logical reasoning. This approach is useful in game of chances like coin tosses, dice game and card games.

The classical approach of probability has serious limitations when we try to apply it to the less orderly decision problems which we encounter in real life situations.

The classical approach is based on the assumption of equally likely events, which really exist in real life situations. These situations are disorderly and unlikely e.g. a man jumps from the multi-story building. The chances of survival and death are not equality likely. The chance of defective product produced by a machine, chances of survival after the age 80, probability that a tyre will burst in less than 20,000 km, probability that a electric lamp will burn in less than 1 000 hours etc. are not equally likely events. Classical approach does not give answer in these situations.

7.4. (2) Relative Frequency Approach:

The classical approach does not apply as we deviate from the game of chances. Moreover, the classical approach fails to give answer in these situations for instance, probability of producing a defective part by a machine, it is not half. When a drug is administered to a patient, probability is not $\frac{1}{2}$ that he will cure and $\frac{1}{2}$ he will not cure. In these situations, we have to conduct a experiment and make a large number of trials only then probability can be ascertained. If in a random experiment, N trails are conducted and event A occurs in N_A trials, then we can say

$$\text{Probability of occurrence of A in N trials} = \frac{N_A}{N}$$

Now, if N is very large, we would not expect much variation in the proportion N_A/N as N increases, i.e. the proportion of occurrence of A will remain approximately constant.

If a coin is tossed 10 time, we may get 7 heads and 3 tails. The probability of head in a single trial is 0.7 and of tail is 0.3. If the coin is tossed 100 times, one may get 60 heads and probability of head is 0.60. If this experiment is repeated large number of times, the probability of head approaches to 0.5. Thus, the relative frequency approach is defined as $P\ (A) = n\ \ lt\ \ \frac{N_A}{N}$.

In practice, we obtain the value of P(A) on the basis of large number of observations. The empirical probability of P(A) can never be obtained because of limit, one can have only a good approximation of P(A) by making a sufficiently large. In this case, one can calculate the probability only after completing the experiment, that is why it is called posteriori probability or empirical probability.

7.4. (5) Subjective Probability Approach:

Subjective probability are based on the beliefs of the person making probability assessments. Thus it may differ from person to person or from time for the same person. I may presently believe that probability of effectiveness of a drug to cure a particular diseases is 0.80 but someone else may believe it is not more than 0.60. Frequent betting in car race, horse race, elections, stock exchange etc. is a clear indication of different probability assessments made by individuals. In stock exchange, one set of investors thinks, the price of a particular share will go down and sell it, whereas other investors thinks it will rise and buy it.

In subjective probability assessment individual makes estimates on the basis of past experience, and education. Sometimes it might be hunches or intuition. The individual uses whatever evidence is available and temper this with his personal feelings about the situation. Suppose a company in launching a new product and top management makes assessments about its success rate. A company is commissioning a nuclear power plant on a site where there is evidence of a geological fault. Management is interested in knowing the probability of a major nuclear accident at this location. In these situation neither classical nor relative frequency approach helps in calculating the probabilities.

7.5. Commonly Used Definition and Rules:

Before discussing the probability rules. It is necessary to know the various types of events. The events can be equally likely, mutually exclusive, dependent and independent, simple and compound event etc.

7.5. (1) Equally likely events :

Events are said to equally likely, when the chances of occurrences of all the outcomes are equal and none of the outcome is more or less, likely than the other, for example, coin tossing, dice rolling and card games. If the sample space "S" consists of n equally likely basic outcomes 0_1, 0_2, 0_3., and on than each of these outcomes have a probability equal to 1/n.

7.5. (2) Mutually Exclusive Events:

Two event are mutually exclusive or incompatible when the both cannot happen in a single or if one event occur then other cannot occur in a single trial.

In coin tossing experiment, if head occurs then tail cannot occur and vice versa. A employee will be a member of trade union or not. A student selected will be a male or female. Both the events can't happen simultaneously in a single trial. It can be represented through a Venn diagrammed.

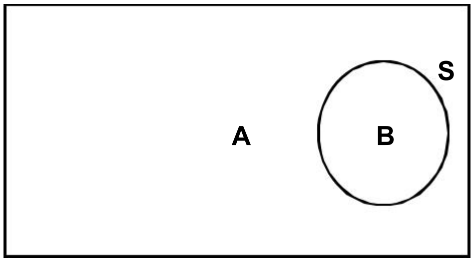


Fig. 1

The entire sample space is represented by a rectangle. If two event are not mutually exclusive they are represented as follows:

In case of mutually exclusive events both the event A and B cannot occur simultaneously hence.

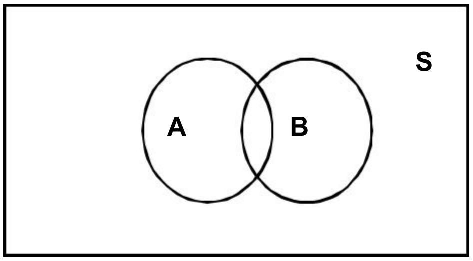


Fig. II

P (occurrence of A and B) = O i.e. impossible

That's why mutually exclusive events are connected by the words either .... or". Event either A or B can occur simultaneously in a single trial as depicted by the shaded area in the Venn diagram (Fig. II).

7.5 (3) Independent and Dependent Events:

Two events is said to be independent, when the outcome of one event does not affect, and is not affected by the another event. Suppose a coin is tossed twice then the result of the second trial is not influenced by the result of the first trial. Similarly, the result of the third throw is not affected or influenced by proceeding outcomes.

From a pack of cards, one card is drawn. The probability of drawing a king is 4/52. Suppose the card drawn is king and again replaced in the pack. The probability of drawing a king in the second draw will be again 4/52. This shows that probability of successive drawn would in no way be affected by the preceding draw.

On the other hand, if the card is not replaced in the pack than probability of the successive draws will be influenced. For example, if king is drawn in the first draw, than the probability of king in the second draw will be 3/ 51. If a king is not drawn in the first draw, then probability will be 4/51. These are called dependent events i.e. outcome of one affect and is affected by the other out-time.

7.5. (4) Simple and Compound Events :

Simple event is a single possible outcome of an experiment whereas compound event is a combination or aggregate of simple events. The occurrence of two or more simple events simultaneously is called compound events. For example probability of drawing three red cards in the first draw and three black cards in the second draw.

7.6. Probability Rules:

Decision makers whose are probabilities are concerned with two conditions:

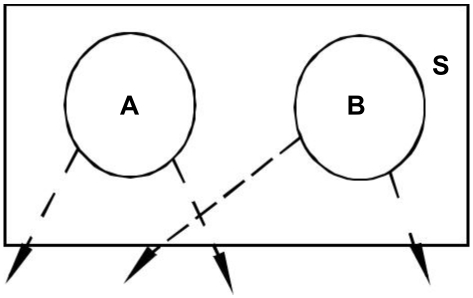
%4 uThe case where one event or another event will occur i.e. probability of either A or event B.

%4 u The case where two or more events will occur simultaneously i.e. probability of occurrence of event A and B.

7.6. (1) Additional Rule Mutually Exclusive Events:

Probability of occurrence of event A or B when both are mutually exclusive is calculated as follows: P (A or B) = P (A) + P (B)

The addition rule can be explained by Venn diagram, where the area of two circles together is the sum of the areas of the two circles.



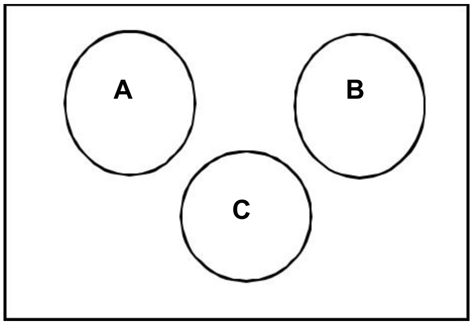
P (A   B) = P (A or B) = P (A) + P (B)

Suppose five candidate A, B, C, D and E applied for a job in a company and there is only one position. The probability that A will be selected is 1/5. Now, we wanted to know the probability either A or B will be selected. Then;

P (A or B) = P (A) + p (B) = 1/5 + 1/5 = 2/5 as both are mutually exclusive events. Both can't be selected simultaneously for one post.

The additional rule can be generalized to more than just two events. The general rule which is clear form the Venn diagram is;

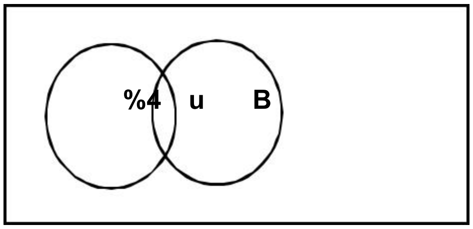
P (A or B or C) = P (A) + P(B) + P(C)



7.6. (2) Additional Rule for Not Mutually Exclusive Events :

There can be situation when two event are not mutually exclusive events, it is possible for both to occur. In these situations, we have to modify our addition rule. Suppose a card is drawn form a pack of cards and we want to calculate the probability of drawing a ace or a heart. In this case both the events ace and heart can occur together i.e. ace of hearts. The ace and heart are not mutually exclusive events. If a superior tells an employee that you will be promoted if you fetch a contract with firm. A or with firm B. Obviously, the employee will be promoted if he gets a contract from both the firms.

It can be explained with the help of a Venn diagram as follows:



If we add the area of two circles A and B, we double count the shaded area so we must subtract it, to ensure it is counted once.

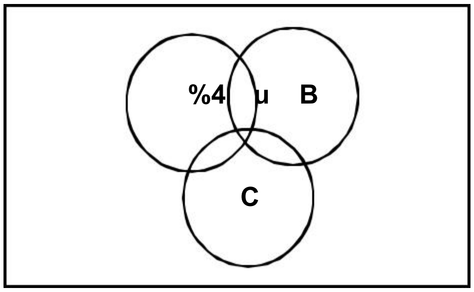
P(A or B) = P(A) + (B) - P (A and B)

The probability of drawing either an ace or a heart card is,

P (ace)	= 4/52
P (heart)	= 13/52
P (ace or heart)	= 4/52+13/52-1/52 = 16/52

The reason for subtracting is quite evident as P (ace of hearts) is counted twice, once in P (ace) and once in P (hearts). This is called double counting error. The horse has eight legs (two in front, two at left two at back and at right) is a double counting error.

The addition rule for more than two events when they are not mutually exclusive events needs modification as is evident from the Venn diagram, the area which is counted twice, should be subtracted.



P(A or B or C) = P(A) + P(B) + P(C) - P(A and B) - P (B and C) - P (C and A) + P (A and B and C)

7.6.(3) Multiplication Rule for Independent Events :

The probability of two or more independent events occurring together or in succession is the product if their probabilities.

P(AB) = P(A and B) = P(A).P(B)

P(AB) or P (A and B) stands from the probability of event A and B occurring together or in succession, it is also known as joint probability.

What is probability of getting a head in the first toss and head in the second toss.

P(H_1 and H_2) = P(H_1) P(H_2)

The multiplication rule can be extended to three or more independent events occuring together P(A, B and C) = P(A). P(B). P(C)

The probability to three heads in the three successive tosses is; P(H_1, H_2 and H_3) = P(H_1). P(H_2), P(H_2)

$\frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}$ = 1/8 = 0.125

Exercises

%4 u A candidate is called for interview for three-posts. For the first post there are 3 candidates, the second there are four and for the third there are two candidates. What are the chance of getting at least one post?

(Hint; (CA 1980)

P (rejection in one post) = 2/3

P (rejection in second post) = 3/4

P (Rejection in third post) = $\frac{1}{2}$

P (A and B and C) = P(A) P(B). P(C)

P (Rejection in one post and second post and third post) = 2/3 3/4 $\frac{1}{2}$ = $\frac{1}{4}$

P (of selection in at least one post) = 1 -- $\frac{1}{4}$ = 0.75]

%4 u A piece of electronic equipment has two essential parts, A and B. In the past, Part A failed 40% of the time and part B fails 50% of the time. Part A and B operate independently. Assume both must operate to enable the equipment to function. What is the probability that the equipment will function ?

[Hint: (M.B. A. Delhi, 1984)

P(A failed) = 4/10

P(B failed) = 5/10

P (A not failed and B not failed = 6/10 5/10 = 0.30

For clearly, I am solving in another way equipment will not function if A failed and B does not fail, the probability is 4/10 5/10 = 0.20.

A does not fail and B failed, the probability is 6/10 5/10 = 0.30.

A failed and B failed, the probability is 4/10 5/10 = 0.20.

The probability the equipment will not function is 0.20 + 0.30 + 0.20 = 0.70.

The probability of the equipment will function = 1 0.70 = 0.30]

%4 u From a set of 25 cards numbered 1, 2, 3....25, one card is drawn at random. What is the probability that a card drowns bears a number which is divisible by

a. 2, b. 3, c. 2 or 3

[Hint:

%4 u The number divisible by 2 are 2, 4, 6....24 i.e. 7/12 numbers, the probability of selecting a number divisible by 2 is 12/25.

%4 u The number divisible by 3 are 3, 6, 9, 12....24 i.e. numbers, the probability of selecting a number divisible by 3 is 8.25.

%4 u The number is divisible by 2 or 3 is. The number, divisible by 2 or 12 and by 3 are 8. But this is not a mutually exclusive event as there are numbers divisible by 2 and 3 both like 6,12, 18, 24. These are counted twice. Hence the probability of selecting a number divisible by 2 or 3 is P(2 or 3) = (divisible by 2) + P (divisible by

P (divisible by 2 and 3) = 12/25 + 8/25 4/25 = 16/25]

%4 u Three ships X, Y and Z sail form England to India chances in favour of their arriving safely are 2 : 3,

3 : 5 and 5 : 7. Find the chances.

%4 u they all arrive safely

%4 u at least one arrives safely

[Hint: P (all arrives safely) = P (X arrive safely)

(Y arrives safely). P(Z arrives safely)

(All events are independent in nature)]

%4 u Two-fighter planes A and B took a flight for a special mission i.e. to destroy a bridge. The past record reveals the probability of A's hitting the target is 3/4 and B's 2/3. Find the probability that the mission is achieved when both fire the shot.

Hint: P (A's hitting the target) = 3/4 P

(B's hitting the target) = 2/3

The mission is achieved when either A hits the target and B miss it of A miss the target and B hits the target of both hits the target

P (A hits the target and B miss it) = 3/4 1/3 = 3/12

P (A miss the target and B hit it) = $\frac{1}{4}$ 2/3 = 2/12

P (both hit the target) = 3/4 2/3 = 6/12

Probability (the bridge will be destroyed) 3/12 + 2/12 + 6/12 =

11/12 It can be solved in another way also.

Firstly find the probability both miss the target P (A and B both miss the target) =

1/4 1/3 = 1/12 P (the bridge is destroyed) = 1-1/12 = 11/12

%4 u The problem is given to four students A,B,C, and D. The chances of their solving the problem are 1/2, 1/3, 1/4,

1/5. Find the probability that the problem will be solved.

[Hint: Probability (none could solve the problem) = 1/2 2/3 3/4 4/5 = 24/120 =

1/5. Probability (problem is solved) = 1 - 1/5 = 4/5

%4 u An article manufactured by a company consist of two parts A and B. In the process of manufacture of Part A, 9 out of 100 are likely to be defective. Similarly, 5 out of 100 are likely to be defective in the manufacture of Part B. Calculate the probability that the assembled part will not be defective.

[Hint: (CA1976)

P (A is non-defective) = 91/100

P (B is non-defective) = 95/100

P (A and B both are non-defective)

= P (A is non-defective) P

(B is non-defective)

= 91/100 95/100 = 0.8654

%4 u The odds against student X solving the problem are 8 : 6 and odds in favour of student Y solving the same problem are 14 : 16.

(i) What are the chances that the problem will be solved of they both try independently of each other ?

(ii) What is the probability that neither solves the problem?

(Hint: Probability (both try and fails to solve it) = P (A fails) P (B fails) = 8/1416/30 = 32/105.

Probability (the problem is solved) = 1- (probability both fails) = 1-32/105 = 73/105].

%4 u The odds that A speaks the truth is 3 : 2 and odds that B speak truth is 5 : 3 in what percentage of cases are they likely to contradict each other on a identical point.

[Hint: (MBA D.U. 1983)

Find the probability A speaks truth and B tells lie and probability of A tells lie and B speaks truth only then contradiction arrives. [Ans. 19/40]

%4 u Suppose from the M. Com. examination, the following information about the result were collected.

%4 u percent failed in Course I

%4 u percent failed in course II

%4 u percent failed in course III

%4 u percent failed in both Course I and II

5 percent failed in both Course II and III

7 percent failed in both course II and I

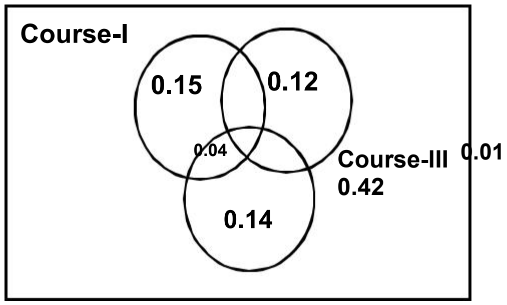
4 percent failed in all the three course.

Suppose a student is selected at random from the class, what is the probability that the randomly selected student is passing in all the three courses.

[Hint: The Venn diagram can Se applied in solving these types of problems.

Probability of failing in all the three courses is 0.04. Probability in failing in course I and II is 0.12.

Hence the probability of failing in course I and II, but not in III is 12-0.04 = 0.08. Similarly the probabilities for others can be computed.



Probability of failing in course II and III but not in I is 0.05-0.04 = 0.01.

Probability of failing in one course or in two courses or in all the courses is equal to ; = 0.15 + 0.08 + 0.04 + 0.03 + 0.14 + 0.01+0.12 = 0.57.

Hence, the probability a candidate is passing in all the three courses is 1 - 0.57 = 0.43.

7.6. (4) Multiplication Rule for Dependent Events:

Multiplication rule is not applicable when the events are dependent. Two event A and B are dependent when event A will occur if event B has already happened. This is known as conditional probability. The conditional probability is written as P (A/B) i.e. probability of event A given that event B has occurred. For example, "I will go out if it does not rain" can be stated in terms of conditional probability as;

P (I will go/no rain)

The conditional probability that event A will happen given that event B happens equals to joint probability of A and B divided by the marginal probability of B (the condition). Mathematically, it is;

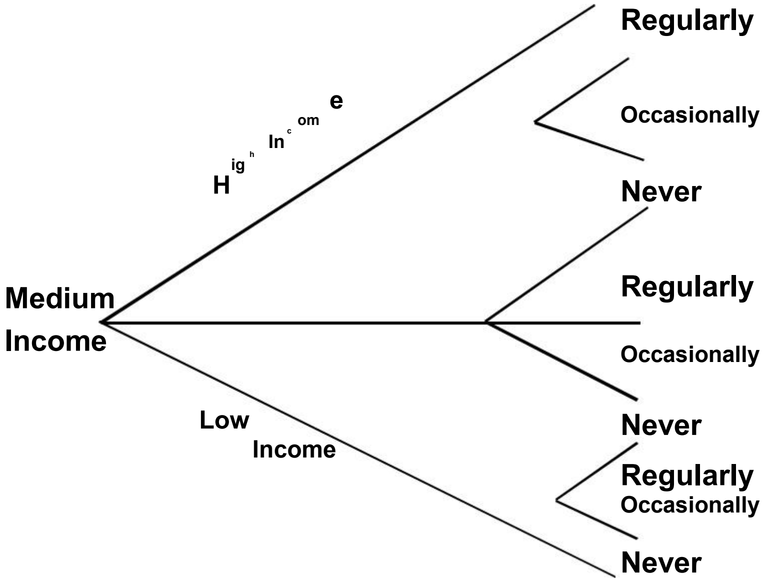
$$P(A/B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(A/B) = \frac{P(A \text{ and } B)}{P(A)}$$

If we solve this for P (A and B) by cross multiplication, we have the formula for joint probability under conditions of statistical dependence:

P (A and B = P (B/A). P (A) or = P (A/B) P(B)

In an example, advertiser wants to know the relationship viewing frequency of a particular TV programme and characteristic of the family i.e. income. The families may be divided into whether they regularly, occasionally or never watch the particular TV programme and also income wise i.e. high, medium and low. Then nine possible outcomes can occur.



Suppose, the number of frequencies for Television viewing-income are as follows:

	High income	Middle income	Low income	Total
Regularly	60	110	50	220
Occasionally	50	140	160	350
Never	60	160	210	430
Total	170	410	420	1000

The above contingency table; can be converted into probabilities, which is as follows:

	High income	Middle income	Low income	Total
Regularly	0.06	0.11	0.05	0.22
Occasionally	0.05	0.14	0.16	0.35
Never	0.06	0.16	0.21	0.43
Total0.17	0.41	0.42	1.00	

On this basis, the probability that a family chosen at random from a population has low income and occasionally watches TV programme is;

P (occasionally and low income) = 0.16

Similarly, probability that a family chosen at random regularly watches the TV programme is;

P (regularly watch) = P (regularly watch and High income) + P (regularly watches and middle income) + P (regularly watches and low income).

The additive rule is applied as high income, middle income and low income are mutually exclusive events.

P (regularly watch) = 0.06 + 0.11 + 0.05 = 0.22 example the probability that a randomly chosen family occasionally watches the show given its income is low, will be

P (occasionally watch Low income) $\dfrac{\text{P(occasionally watch low income)}}{\text{P(low income)}}$ $\dfrac{0.16}{0.42}$

Similarly P (occasionally watch high income) $\dfrac{\text{P(occasionally watch and middle income)}\ \ 0.05}{\text{P(High income)}0.17}$

P (occasionally watch middle income)

	High income	Middle income	Low income
Regularly	$\frac{0.06}{0.17}$ 0.35	$\frac{0.11}{0.41}$ 0.27	$\frac{0.05}{0.42}$ 0.12
Occasionally	$\frac{0.05}{0.17}$ 0.29	$\frac{0.14}{0.41}$ 0.34	$\frac{0.16}{0.42}$ 0.38
Never	$\frac{0.06}{0.17}$ 0.35	$\frac{0.16}{0.41}$ 0.39	$\frac{0.21}{0.42}$ 0.50

Similarly, P (occasionally watch) = 0.55

P (Never watch) = 0.43

P (High income = 0.17

P (middle income) = 0.41

P (low (income) = 0.42

The conditional probabilities can be obtained from the above contingency table. For

$$\frac{P(\text{occasionally watch and middle income})\quad 0.14}{P(\text{High income})0.41}$$

Similarly, the other conditional probabilities of watching TV programme given income level can be ascertained for regularly and never watching TV programme. It will be conditional Probabilities of Viewing Frequencies of TV Programme given income level.

	High income	Middle income	Low income
Regularly	$\frac{0.06}{0.22}$ 0.27	$\frac{0.11}{0.22}$ 0.50	$\frac{0.05}{0.22}$ 0.23
Occasionally	$\frac{0.05}{0.35}$ 0.14	$\frac{0.14}{0.35}$ 0.40	$\frac{0.16}{0.35}$ 0.40
Never	$\frac{0.06}{0.43}$ 0.14	$\frac{0.16}{0.43}$ 0.37	$\frac{0.21}{0.43}$ 0.49

The conditional probability of income levels given a viewing frequency of TV programme can be calculated in the same manner. It will be.

Conditional probabilities of income levels given viewing frequencies.

In this example, for the events never watch and low income, we observe.

P (Never watch and low income; = 0.21 whereas

P (Never watch) = 0.43 and

P (Low income) = 0.42 The product of marginal probabilities is 0.18 which differs form the joint probability i.e. 0.21. Hence, the two events are not statistically independent, in nutshell, we can state.

7.7. Additive law of probability :

The probability of either A or B is equal to

P (A or B) = P (A) + P (B) for mutually exclusive events and

P (A or B) = P (A) + P (B) - P (A and B) for non mutually exclusive events.

7.8. Multiplication law of probability :

The probability of A and B is equal to ;

P (A and B) = P(A). P(B) for independent event and;

P(A and B) = P(A) (B). P (B) = P (B) (A). P(A) for dependent events.

Exercises

11. Consider the following height and weight contingency table

	Short (S)	Medium (M)	Tall(T)	Total
Light (L)	0.03	0.17	0.01	0.21
Medium Light (ML)	0.02	0.33	0.02	0.37
Heavy (H)	0.02	0.20	0.02	0.42
	0.07	0.70	0.23	1.00

Find the probability of;

a. P(L/T) b. P(ML/M)

c. P.(T/L) d. P(M/ML)

e. P(H/T) f. P(T/H)

Hint:

a. $P(L/T) = \frac{P(\text{Land T})}{P(T)} = \frac{0.01}{0.23}$

b. $P(ML/M) = \frac{P(\text{ML and M})}{P(M)} = \frac{0.03}{0.70}$

c. $P(T/L) = \frac{P(\text{and T})}{P(L)} = \frac{0.10}{0.21}$

d. $P(M/ML) = \frac{P(\text{M and L})}{P(L)} = \frac{0.33}{0.37}$

e. $P(H/T) = \frac{P(\text{ H and T})}{P(T)} = \frac{0.20}{0.23}$

f. $P(T/H) = \frac{P(\text{T and H})}{P(H)} = \frac{0.20}{0.42}$

12. Consider the following contingency table.

	Male (M)	Female (F)	Total
Brand Loyal (BL)	0-60	0-05	0-65
Not Brand Loyal (NBL)	0-15	0-20	0-35
Total	0-75	0-25	1-00

Find the Probability of

a. P (BL/F) b. P (F/BL) c. P (M/BL) d. P (BL/M)

[Ans : a. = $\frac{0.05}{0.05}$, b = $\frac{0.05}{0.65}$ c. =, $\frac{0.60}{0.65}$ d. $\frac{0.60}{0.75}$]

%4 u The personnel manager of a large manufacturing firm finds that 15 percent of the firm's employee are junior executive and 25 percent of the firm's employee are MBA's. He also discovers that 5 percent of the firm's employees are both junior executive and MBA's. What is the probability of selecting a junior executive if the selection is confined to MBA's ? (MBA Patiala 1979)

Hint:

P (Junior executive and MBA) = 0.05, P (MBA) = 0.25

P Junior executive/MBA's $\frac{P(\text{Junior executive and MBA})}{P(MBA)}$ $\frac{0.05}{0.25}$ $\frac{1}{5}$ 0.20 .

%4 u A company learned that inventory shortage were associated with a loss of goodwill with a probability 0.10. The company also knew that a loss of goodwill from all sources occured with a probability of 0.15. What is the probability of an inventory shortage, given a loss of goodwill? (MBA, Kurukshetra, 1977)

Hint :

P (Inventory shortage and loss of goodwill) = 0. 1 0

P (loss of goodwill) = 0.1 5

0.10 2

P (Inventory shortage/loss goodwill) – – 0.67.

15. The personnel department of a company has records which show the following analysis of its 200 engineers.

Age	— Bachelor's degree	Master's degree	Total
Under 30	90	10	100
30 to 40	20	30	50
over 40	40	10	50
Total	150	50	200

If an engineer is selected at random :

%4 u The probability he has only a Bachelor's degree.

%4 u The probability he has a Master degree, given that the he is above 40.

%4 u The probability he is under 30, given that he has a Bachelor's degree.

(MBA, Delhi, 1977)

[Ans : a. = $\frac{150}{200}$, b = $\frac{10}{50}$ c. = $\frac{90}{150}$]

7.9. Axiomatic Approach to Probability Theory:

Although there is not much objection against the in logical content of frequency approach for defining probability but there is some inherent weakness or inelegance in the mathematical formalism. In the definition, we note that the frequency ratio is thoroughly an empirical concept, whereas the limit is postulated in a rigorous analytical sense. This combination of empirical and theoretical concept is very inelegant, and naturally leads to mathematical difficulties. Now this problem is not typical of probability theory only, but arises in other branches of mathematics as well e g., the theory of geometry. In geometry, we face the same difficult situation, if we try to define the fundamental entities like a point or a straight line. We may attempt to define a point as the limit of a sequence of chalk dots drawn on the blackboard of gradually decreasing dimensions, which will be similar to the definition of probability. This is, however, not done in modem theories of geometry, in which point, straight line, etc. remains undefined concepts, and we start with a system of axioms which specify the fundamental relations among them. In the theory of probability also, we are ultimately forced to give up the hope of defining probability and take resource to an axiomatic theory in which probability is accepted as an undefined new concept, and only the salient rules for calculation of probabilities are postulated. These rules will, however, be chosen from the previous theories with necessary modification for operation convenience.

7.10. Axioms of Probability:

The axiomatic approach to probability was proposed by A.N. Kolmogorov in the year 1983. When this approach is followed, no precise definition of probability is given, rather we give certain axioms or postulates on which probability calculations are based.

Let E be a random experiment described by the sample space S (set of all possible outcome), and A any event connected with any event i.e. A S. The portability of the event A is a number associated with A to be denoted by P (A), such that the following axioms are satisfied:

I. Axioms of Postiveness :

To each event A (outcome), there can be assigned to non negative real number P [A], i.e. a number such that 0 P[A].

This number is called the probability of the event A. While we assumed that o P(A) I, we require here only that o P(A). Although the earlier hypothesis seems to be the natural one in view of our desire to have probability model relative frequency, it can be shown that P(A) I follows as a conclusion from our present axioms and hence that it does not have to be made as a priori hypothesis.

II. Axioms of Certainty :

The probability of a certain event P(S) = I.

This axioms stated that the probability of the sure, or certain event is unity. It should be realized, however that the converse statement, i.e., if P (A) = I, then A = S, does not necessarily hold true.

III. Axioms of Union :

If A_1, A_2, A_3, ... be a finite or infinite sequence of pairuise mutually exclusive events i.e.

$A_i \quad A_j = \quad \text{for } i \quad j$

(i, j, = 1, 2, 3, ...)

Then $P(A_1 + A_2 + A_3 + ...) = P(A_1) + P(A_2) + P(A_3) + ...$

$$\underset{i=I}{\text{or } PA_i} \qquad \underset{i=I}{P(A_i)}$$

A moments reflection upon the last two axioms should show that we have defined how to determine the probability of a union of events only for those cases in which the events under consideration are mutually exclusive or disjoint. It is for this reason that whenever we are faced in the future with a problem in which we are given an arbitrary collection to events, we will generally have to construct from them a new collection of disjoint sets.

Now starting from the above axioms, we can logically build up the mathematical structure of the theory of probability. But in order that such a theory may also be meaningful from the point of view or practical applications, we must have to postulate the basic rule for connection the ideal number probabilities with experience. This rule, not included in the axioms, consist in the following interpretation (not frequency definition) of probability. If the random experiment E is repeated a large number of times under identical or uniform conditions, the frequency ratio of any event will be approximately equal to its probability i.e. $P(A) \simeq f(A)$ so that f (A) can be taken to be an experimentally measured value of the idealized number P (A), and large is the sequence of repetitions of E more accurate is the measured value.

7.11. Baye’s Theorem:

We shall be now discussing a very powerful statistical methods known as Baye’s Theorem for evaluating new information and revising our prior estimates (based upon limited information only) of the probability. If correctly used, it makes it unnecessary together masses of data over long periods of time in order to make good decisions based on probabilities.

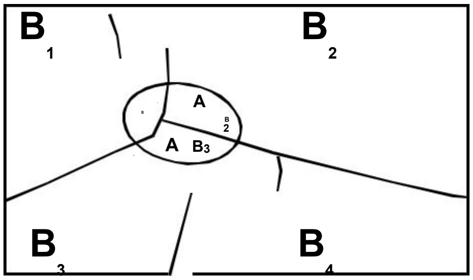
For example, Indian Hockey Team coach find that the selected players lack stamina or some of them are prone to injury or lack dedication or can not do practice for long periods. He has to change his strategy for better results. A similar situation occurs in business also. If a manager of a firm finds that most of the purple ski jackets thought would sell well are hanging on the rack, he most revise his prior probabilities and order different colour combination to have a sale. Take an example of a person who wants to invest his saving in shares/debenture/fixed deposit of bank to have a maximum income. His investment decision will also depend on current and past situation of the market, interest rate offered by the company or bank and also about the future trend. In the above cases, certain probabilities have to be altered after tie people involved got additional information. The probability theory is of great value in managerial decision making.

So it is not easy to compute conditional probabilities directly from the conditional probability formula provided that P (B) is not zero. A convenient formula that gives the relationship among various conditional probabilities is Baye’s theorem. Before Baye’s theorem is stated and proved let us define Partition of a sample space and state the total probability law:

$$P(A/B) \frac{P(A \quad B)}{P(B)}$$

Partition of a sample space :

A partition of the sample space is defined as follows: If B_1, B_2,........, B_K are mutually exclusive (disjoint) subsets of S and B_1 U B_2 U B_K = S (exhaustive) then these subsets are said to form a partition of S. When the experiment is performed, one and only one of the events B_i , occurs if we have partition of S.



In general, if k events B_i (i = 1, 2,......., k), form a partition and A is an arbitrary event w.r.t. S, then we may write A = (A B_1) (A B_2) .... (A B_K)

so that P(A) = P(A B_1) P (A B_2) + ... + P (A B_K)

[Axiom of Union]

Since the events (A B_i) are pair-wise mutually exclusive (see figure for k = 4). It does not matter that (A B_i) = for some or all of the i since P ( ) = 0.

Result of Total Probability Law :

Theorem : If B_1, B_2, ... B_k represents a partition of S and A is an arbitary event on S, then total probability of A is given by

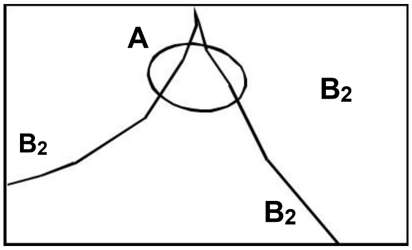
P(A) = P(B_1 A) P (B_2 A) + ... + P (B_K A)

%4 uP(B_1). P(A/B_1 P (B_2). P(A/B_2P(B_K P(A/B_K P(A
B) P(A) P (B/A)

The result of the above theorem is very useful, or there are numerous practical situations in which P(A) cannot be completed directly. However, with the information that B_i has occurred, it is possible to evaluate P (A/B_1) and this determine P(A) when the values P(B_i) are obtained.

Ex. : In 1984. there will be three candidates for the position of principal Mr. Chatterji, Mr. Ayangar and Dr. Singh-whose chances of getting the appointment are in the proportion 4:2:3 respect. The probability that Mr. Chatterji if selected will introduce co-education in the college is 0.3. The probability of Mr. Ayangar and Dr. Singh doing the same are respectively 0 5 & 0.8. What is the probability that there will be co-education on in the college in 1984.

Sol : Let the events and probabilities be defined as follows:



%4 u : Introduction of co-education.

B_1 : Mr. Chatterji is selected as principal.

B_2 : Mr. Ayangar is selected as principal.

B_3 : Dr. Singh is selected as principal.

Then P (B_1) = $\frac{4}{9}$,P(B_2) = $\frac{2}{9}$,P(B_3), = $\frac{3}{9}$.

P (A/B_1) = 3/10, P(A/B_2) = 5/10 and P(A/B_3) = 8/10

P (A) = P [(A B_1) (A B_2) (A B_3)]

P (A) = P (A B_1) [Axiom of union]

+P(A B_3)

$P(B_1 \quad A,) \;\; P(B_2 \quad A) + \ldots\ldots + P(B_3 \quad A)$

%4 u $P(B_1)\,P(A/B_1) + P(B_2)\,P(A/B_2) + P(B_3)\,P(A/B_3)$

$$9^4 , \; 9^3 \; 9^2 , \; 10^5 \; 9^3 , \; 10^8 \quad 42^{23}$$

Another important result of the total probability law is known as 'Baye's Theorem.

Theorem (Baye's Th.) :

If $B_1, B_2, \ldots\ldots B_n$ are mutually disjoint exhaustive events with $P(B_i)$ 0 (i = 1, 2,.......n) then for any arbitrary

events A which is a subject of $\overset{n}{\underset{i=I}{}} B_1$ such that P (A) > 0, we have

$$P(B_i/A) = \frac{P(B)i/P(A/B_i)}{\overset{n}{\underset{i=I}{P}}(B_i)P(A/B_i)}$$

Proof: Since A $\overset{n}{\underset{i=I}{}} B_i$, we have

$$A = A \quad (\overset{n}{\underset{i=I}{}} B_i) = \overset{n}{\underset{i=I}{}} (A \quad B_i)$$

Since $(A \quad B_i) \quad B_i$, by distributive law

(i = 1, 2,....... n) $(A \quad B \quad C) = (A \quad (A \quad C)$

are mutually disjoints events, we have by addition theorem of probability (or Axiom of union)

$$P(A) = P\overset{n}{\underset{i=I}{}}(A \quad B_i) \quad \overset{n}{\underset{i=I}{}} P(A \quad B_i) \quad \overset{n}{\underset{i=I}{}} P(B_i \quad A) \quad \overset{n}{\underset{i=I}{P}}(B_i)P(A/B_i) \qquad (1)$$

= [(P (A B) = P (A) P (B/A) = P (B) P (A/B)] Also, we have P (A B$_i$) P (A) P (B$_i$/A)

$$P(B_i/A) = \frac{P(A\,B_i)}{P(A)} \quad \frac{P(B_i).P(A/B_i)}{\overset{n}{\underset{i=I}{}}P(B_i)P(A/B_i)} \qquad \text{from = n (1)}$$

Remarks:

%4 u The probabilities $P(B_1), P(B_2), \ldots\ldots\ldots, P(B_n)$ are termed as the a prior probabilitie's because they exist before we gain any information from the experiment itself.

%4 u The probabilities P (A/B$_i$). ii = I, 2,.......,n are called 'likelihood' because they indicate how likely the event A under consideration is to occur, given each and every a prior probability.

%4 u The probabilities P (B$_i$/A), i = 1,2,.......,n are called 'posteriori probabilities' because they are determined after the results if the experiment are known.

Some interesting points about Baye's Theorem are:

%4 u Though it deals with a conditional probabilities, its interpretation is different from that of the general conditional probability theorem. The general conditional probability theorem asks "What is the probability of the sample or experimental result given the state value? When is Baye's theorem asks : What is the probability of the state value given the sample or experimental result."

%4 u When we talk of Baye's Theorem, different decision makes may assign different probabilities to the same set or states of nature. Also we may conduct a new experiment by using posterior probabilities. As we proceed with repeated experiments, evidence accumulates and modifies the initial prior probabilities, thereby modifying the intensity of a decision-makers belief in various

states of nature. In other words, the more evidence we accumulate, the less important are the prior probabilities.

%4 u The nation of ‘prior’ and ‘posterior’ in Baye’s theorem are relative to a given sample outcome. That is, if a posterior distribution has been determined from a particular sample, this posterior distribution would be considered the prior distribution relative to a new sample.

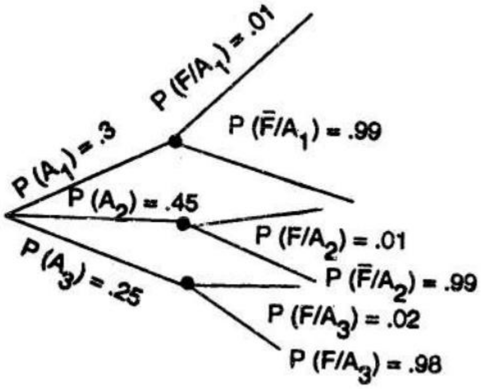
Ex : A Shopekeeper buys a particular kind of light bulb from there manufactures A_1, A_2 and A_3. He buys 30% of his stock from A_1, 45% from A_2 and 25% from A_3. In the past the he has found that 2% of A_3’s bulb are faulty whereas only 1% of A_1’s and A_2’s are. Suppose that he chooses a bulb and finds it is faulty. What is the probability that it was one of A_3’s bulbs ?

Sol.: Here P (A_1), P (A_2) and P (A_3) are prior probability because they exist before we gain any information from the experiment itself. If he picks a bulb at random P (A_1) = 0.3, P (A_2) = 0.45 & P (A_3)= 0.25.

The probability of faulty bulb will be greater than P (A_3) = 0.25. since A_3 produces a greater proportion of faulty bulbs than A_1 & A_2. If F is the event that the bulb is faulty, then P (A_3/F) is the probability that we require. We have

P (F/A_1) = .01, P (F/A_2) = .01, P (F/A_3)
= .02 (are likelihood probability).

This information is shown on a tree diagram.



$$P\,(A_3/F) = \frac{P(A_3)\,P\,(F/A_3)}{P(A_1)\,P\,(F/A_1) + P(A_2)\,P\,(F/A_2) + P(A_3)\,P\,(F/A_3)}$$

by using Barye’s Th. $P(B_i/A) = \frac{P(B_i)\,P\,(A/B_I)}{\sum_{i=I}^{n} P(B_i)P(A/B_i)}$ $\frac{.25 \quad .02}{.3 \quad .01 \quad .45 \quad .01 \quad .25 \quad .02}$ 0.4.

Similarly, if we wish to know the probability that the faulty bulb was supplied by A_1 or A_2, we have

$$P\,(A_3/F) = \frac{P(A_1)\,P\,(F/A_1)}{P(A_1)\quad PP\,(F/A_1) + P(A_2)\,P\,(F/A_2) + P(A_3)\,P\,(F/A_3)}$$

$$\frac{.3 \quad .01}{.3 \quad .01 \quad .45 \quad .01 \quad .25 \quad .02} 0.24.$$

and since A_1, A_2 and A_3 are mutually exclusive P (A_2/F) = 1 - 0.4 - 0.240 = .36.

P (A_2/F) = 1 - 0.4 - 0.240 = .36.

Ex.: The contents of urn I, II and III are as follows :

1 white, 2 black and 3 red balls,

2 white, 1 black and 1 red ball and

4 white, 5 black and 3 red balls.

One urn is chosen at random and two balls are drawn. They happen to be white and red. What is the probability they come from I, II or III ?

Let B_1, B_2 and B_3 denote the events that the urn I, II and III is chosen, respectively, and let A be the event that the two balls taken from the selected urn are white and red Then

$$P(B_1)\ \ P(B_2) = P(B_3) = {}^{1}\cancel{3}$$

$$P(A/\ B_1)\ \ \frac{{}^{1}C_{1}\ \ {}^{3}C_{1}}{{}^{6}C_{2}} = \frac{1}{5} \qquad {}^{n}C_{r}\ \ \frac{n}{|\,n\text{-}r\ \ r}$$

$$P(A/\ B_3)\ \ \frac{{}^{4}C_{1}\ \ {}^{3}C_{1}}{{}^{12}C_{2}}$$

Hence $$P(B_2\,/A) = \frac{P(B_2)\,P\,(A/B_2)}{{}^{3}_{i=l}P(B_i)P(A/B_i)} \qquad \frac{1/3\ 1/3}{1/3\ 1/5\ \ 1/3\ 1/3\ \ 1/3\ \ 2/11}\,55\,/118.$$

Similarly, $$P(B_3\,/A) = \frac{1/3\ \ 1/3}{1/3\ \ 1/5\ \ 1/3\ 1/3\ \ 1/3\ \ 2/11}$$

$$P(B_1\,/A) = 1\quad \frac{55}{118}\ \ \frac{30}{118}\ \ \frac{30}{118}\ .$$

Ex.: A company has two plants to manufacture scooters, Plant I manufacture 70% of the scooters and Plant II manufacture 30%. At plant 1,80% of the scooters are rated of standard quality and at plant II, 90% of scooters are rated of standard quality. A scooter up at random and is found to be of standard quality. What is chance that it has come from plant I ?

Sol. : Let B be the event of picking a standard quality scooter and let A_1 and A_2 be the events of picking from plant I and plant n respectively. We want to calculate P (A_1/B)

$$P(A_1\,/B) = \frac{P(A_1)\ P\,(B/A_i)}{{}^{2}_{i=l}P(A_i)P(B/A_i)}$$

P (A_1) = 2/3, P(A_2) = 1/3 P(B/A_2) = 2/100 = .02 and P(B/A_2) = 0.1.

$$P(A_1\,/B) = \frac{2/3\ \ 0.2}{2/3\ \ .02\ \ 1/3\ \ .01}\ .80.$$ **is the required probability.**

Self Assessment Fill in the Blanks:

%4 uThe concept of probability originated from the analysis of the in the 17th century.

%4 uThe theory of probability is a study of Experiments.

%4 uA phenomenon or an experiment which can result into more than one possible outcome, is called a or statistical experiment.:

%4 uMathematical definition of probability, was given by.

%4 uClassical definition is also Known as definition of probability.

%4 uTwo or more outcomes of an experiment are said to be if the occurrence of one of them precludes the occurrence of all others in the same trial t.e. they cannot occur jointly.

%4 uA .... is an arrangement of a given set of objects in a definite order.

%4 u When no attention is given to the order of arrangement of the selected objects, we get a....................

Self Check exercise

%4 uDefine the term 'probability' by The Classical Approach.

%4 uDiscuss equally likely events.

%4 uDiscuss Independent and Dependent events.

%4 uDiscuss the axiomatic approach to probability. Exercise

Q. 1. 8% of the bulbs produced by a factory are red and 2% are red and defective. If one bulb is picked up at random, find the probability of its being defective if it is red.

Hint : Let E_1 be the event of the bulb red and E_2 that of it's being defective.

$$P(E_1) = \frac{8}{100}\ .8, P(E_1 \_\_ E_2) = \frac{2}{100}\ .02$$

Prob. of picking up a defective bulb if it is red = $P(E_2 / E_1) \frac{P(E_2 E_1)}{P(E_1)}$ Ans : 1/4.

Q. 2. A bag A contains 2 white and 3 red balls and a bag B contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and is found to be red. Find the probability that it was drawn from bag B.

Hint: Let E, be the event that the ball is drawn from the bag A, and E_2 be the event that it is drawn from the bag B and E be the event that the drawn ball is red.

$P(E_1) = 1/2 = P(E)$, $P(E/E_1) = 3/5$, $P(E/E_2) = 5/9$

$$P(E_2/E) = \frac{P(E_2)\ P(E/E_2)}{P(E_1)\ P(E/E_1) + P(E_2)\ P(E/E_2)} \text{ Ans.} \frac{25}{52}.$$

Q.3. A factory has 3 machines, A, B, C producing 1000, 2000 and 3000 bolts per day respt. A produces 1% defective, B 1.5% & C 2% defective. A bolt is checked at random at the end of a day and is found to be defective. What is the probability that it came from machine A ?

Hint: Let E_1, E_2, E_3 be the events that the bolt was chosen from machine A, B & C respt. Let E be the event that it is defective.

$$P(E_1) = \frac{1000}{1000+2000+3000}\ 1/6, P(E_2) = \frac{2000}{6000}\ \frac{2}{6}$$

$$P(E_3) = \frac{3000}{6000}\ \frac{3}{6}\ P(E/E_3) = \frac{1}{100}\ .01, P(E/E_2)\ 0.15, P(E/E_3) = .02$$

$$P(E_1/E) = \frac{P(E_1)\ P(E/E_1)}{{}_{i=l}^{3} P(E_i)P(E/E_i)} \quad \text{Ans : 1}$$

Q. 4. What is the chance that a leap year selected at random will contain 53 Sunday.

Ans : 2/7

Q. 5. Three groups of children contain 3 girls and 1 boy, 2 girls and 2 boys, one girl and 3 boys. One child is selected from each group. Show that the chance that the three selected consist of 1 girl and 2 boys is 13/32.

Hint : " The selection can be made in the following manners

%4 u Boy, boy, girl ;Prob. (1/4) (1/2) (1/4)

%4 u Boy, girl, boy ; Prob. (1/4) (1/2) (3/4)

%4 u Girl, boy, boy ; Porb. (3/4) (1/2) (3/4)

Since these are mutually exclusive events, the regd. prob. is the sum of probabilities of three cases.

Ans : 13/32.

7.12. Summary

The theory of probability has vide applications in different fields of life. Probability is useful in making predictions when there is uncertainty of an event. Probability theory is used in major field like law of Statistical regularities. Law of inertia of large numbers is based on theory of Probability. Various Parametric and non-parametric tests like Z test, t-test, F test etc, are based on theory of Probability. Probability is used in taking economic decision in the situation of risk and uncertainty by the Sales managers, Production Managers etc. Probability is also useful in various scientific investigations.

7.13 Glossary:

Combination: When no attention is given to the order of arrangement of the selected objects, we get a combination.

Equally likely outcomes: The outcomes of random experiment are said to be equally likely or equally probable if the occurrence of none of them is expected in preference to others.

Expected Monetary Value: When a random variable is expressed in monetary units, its expected value is often termed as expected monetary value and symbolized by EMV.

Expected Value: Expected value of a constant is the constant itself, i.e., E(o) = b, where bis a constant.

Mutually exclusive outcomes: Two or more outcomes of an experiment are said to be mutually exclusive if the occurrence of one of them preciudes the occurrence of all others in the same trial i.e. they cannot occur jointly.

Permutation: A permutation is an arrangement of a given set of objects in a definite order. Thus composition and order both are important in a permutation.

Priori' definition of probability: If n is the number of equally likely, mutually exclusive and exhaustive outcomes of a random experiment out of which m outcomes are favourable to the occurrence of an event A, then the probability that A occurs, denoted by P(A).

Random phenomenon: A phenomenon or an experiment which can result into more than one possible outcome.

7.14 Answers: Self Assessment

1. games of chance
2. Statistical or Random
3.random phenomenon or random experiment
4. J. Bernoulli
5.'a priori'
6.mutually exclusive
7. Permutation :
8. Combination
9. Refer to section 7.4(1)
10.Refer to section 7.5(1)
11.refer to section 7.5(3)
12.Refer to section 7.9

7.15 Terminal Questions

%4 u Explain the concept of probability.

%4 u Explain the different approaches to Probability.

%4 u A bag contains eight balls, five being red and three white. If a man selects two balls at random from the bag. What is the probability that he will get one ball of each colour.

%4 u State and prove the addition theorem of probability for any two events.

%4 u Explain the concept of 'Inverse Probability' with example.

7.16. Suggested Readings

%4 u Feller W., An Introduction to Probability Theory and its Applications, John Wiley & Sons. Inc New York.

%4 u Hogg. R. V and A Elliott, Probability and Statistical Inference. Mac Millan Publishing Co. Inc…… New York.

%4 u Ronontree, D., Probability, Charles Scribner's Sons, New York,

%4 u New Bold, Paul, Statistics for Business and Economics, Prentice Hall Inc…… Englewood Cliffs.

%4 u Levin Richard Statistics of Management, Prentice Hall Inc….. New Delhi.

%4 u Gupta S.P. Statistical Methods, Sultan Chand & Sons, New Delhi.

*****

Lesson-8
Probability Distribution-Binomial and Poisson

Structure

8.1 Learning Objectives

After studying this lesson, you should be able to understand

%4 u Theoretical or Probability Distribution
%4 u Binomial Distribution
%4 u Properties of Binomial Distribution
%4 u Poisson Distribution
%4 u Properties of Poisson Distribution

8.2. Introduction:

In Statistics, we study different types of distributions. They can be classified in to two headings:-

%4 u Observed Frequency Distribution
%4 u Theoretical or Probability Distribution

Observed Frequency Distribution: It refers to the frequency distributions which are obtained by actual observations or experiments. For example, we may study the marks of the students of a class and classify the data in the form of a frequency distribution as follows

Marks	No of Students
0-10	10
10-20	15
20-30	25
30-40	15
40-50	10

The above example clearly shows that the observed frequency distributions are obtained by grouping data.

Theoretical or Probability Distribution – Theoretical frequency distribution refers to those distributions which are not obtained by actual observations or experiments but are mathematically deduced under certain assumptions.

Such distribution are also called Probability Distribution or Expected Frequency Distribution.

For example, if a coin is tossed we expect that as n increased we shall get close to 50 percent heads and 50 percent heads and 50 percent tails. If a coin is tossed 100 times, we may get 40 heads and 60 tails.

This is our observation. Our expectation is 50 percent heads and 50 percent tails. Amongst Theoretical or expected frequency distributions, the following are more popular:-

%4 u Discrete Probability Distributions

Binomial Distribution
Multinomial Distribution
Negative Binomial Distribution
Poisson Distribution
Hypergeometric Distribution

%4 u Continuous Probability Distribution-

(1) Normal Distribution

In the present chapter, we will discuss the two important discrete random distributions i.e. Binomial and Poisson Distribution.

8.3. Binomial Distribution

One of the most elementary and useful discrete random variable is associated with the coin-tossing experiment. In the experiment, either one coin is tossed 'n' times or 'n' coins are tossed once. The observation 'head or tail' is recorded for each toss. Numerous experiments of practical importance in social sciences, physical sciences, and industry resemble with the coin-tossing experiment. For example, if a house wife is selected at random from a city and asked whether she watches a particular programme of TV, there seem to be only two possible answer 'Yes' and 'no'. Interviewing a single housewife bears a similarity, in many respects, to tossing a single coin because the unit chosen from a production lot can be defective or 'non defective'. A student selected from a university can be male or female, married or unmarried. In a sample survey of voter preference in political election, a voter can have preference or no preference for a particular candidate. Similarly a customer either have a brand preference or no brand preference for a particular product. Firing a projectile at a target may either hit or miss the target. All these experiments are similar to coin tossing experiment resulting in head or tail, yes or no, success or failure.

To standardize the terminology describing these and many other similar process, we call one of the two possible outcomes a success and other as a failure. These names are used only to identify the outcomes and bear no connotation of 'goodness' about the outcome. Such a random phenomenon is known as Bernoulli processes. Since the model involves only two classes of events, it is also known as two point probability model. It was developed by Jacob Bernoulli.

8.4. Assumption of Binomial Distribution:

%4 u The experiment consists of n identical trials.

%4 u Each trial results in one of the two outcomes. For lack of a better nomenclature. One outcome is called success and other as a failure.

%4 u The probability of a success in a single trial is equal to P and of failure is q = 1 - P .

%4 u The trials are independent.

%4 u One is interested in r successes during the n trials.

The two events, success and failure are qualitative in nature. We can convert these qualitative events into numeral ones by assigning the value T to a success and '0' to a failure. Having defined the binomial experiment and its practical applications we will now derive the probability distribution for the random variable. We will obtain the probability distribution for a coin tossing experiment containing n = 1, 2 and 3 coins and generalize the derivation. Let us treat head as a success and tail as a failure.

For n = 1 coin, the possible outcomes are either o success i.e. tail or one success i.e. one head.

Probability Distribution (One coin)

Numbered Head	Coin C_1	Probabilities	P
0	T	q	1/2
1	H	P	1/2

The probability distributions to an experiment consisting of 2 coins will be as follows :

There can be 0, 1 or 2 successes (heads is a success).

No. of heads	Coins		P(r)
	C_1	C_2	
0	T	T	$q.q.=q_2$
1	H	T	P.q
	T	H	q.P=2qP
2	H	H	$P.P.=P_2$

The probability distribution for an experiment consisting of three coins will be as follows :

No. of heads		Coins		P(r)
	C_1	C_2	C_2	
0	T	T	T	$q.q.q=q_{-3}$
	H	T	T	Pqq
1	T	H	T.	$qPq=3q_2P$
	T	T	H	qqP
	T	H	H	qPP
2	H	T	H	$PqP=3qP_2$
	H	H	T	PPq
3	H	H	H	$P.P.P.=P_3$

for one coin, the probability distribution is $(q + P)_2 = q + P$ for two coins, the probability distribution is $(q + P)_2 = q_2 + 2qP + P_2$ for three coins, the probability distribution is $(q + P)_2 = q_2 + 3q_2P + 3qP_2 + P_2$ for n coins the probability distribution i.e. obtaining 0, 1, 2, 3.....n successes can be determined by the successive terms of the binomial expansion $(q + P)_2$ which is

$(q + P)^n = {}^nC_0q^{n+n}\ C_r\ q^{n-1}\ P^{1+n}\ C_2\ q^{n-2}\ P^2 + n\ C_2\ q^{n-3}\ P^{3}$' + ... $n\ Cr\ q^{n-r}\ p^r$........+ $n_{C_n}\ p^r$

These terms can be presented in the probability distribution table as follows :

No. of Heads	Probability
0	${}_nCq_n$ 0
1	${}^nC_1q^{n-1}\ P$
2	${}^nC_2q^{n-2}\ P_2$
3	${}^nC_3q^{n-3}\ P^3$
.	.
.	.
.	.
r	${}^nC_r\ P_{n-r}\ P_r$
.	.
.	.
.	.
n	${}^nC_1\ P^n$

From the above experiment of one, two, three or n coins, we can draw the following inferences.

%4 u Number of items are ‘n + 1’. For example when n = 1 coin these are two, outcomes, when n = 2, these are three terms and so on.

%4 u Sum of powers is equal to ‘n’.

%4 u Order of powers. The power of ‘q’ goes on decreasing by one and of ‘P’ goes on increasing by one.

%4 u All are positive, no negative sign can exist.

%4 u Binomial coefficients can be determined with the help of combination concept or with the help of Pascal triangle which is as follows :

Pascal’s Triangle

Value of N	Binomial Coefficients												
1						1		1					
2					1		2		1				
3				1		3		3		1			
4			1		4		6		4		1		
5		1		5		10		10		5		1	
6	1		6		15		20		15		6		1

From the above table, we can draw an inference that if P = 0.5 then binomial distribution is symmetrical and if P %4 u 0.5, it is skewed to the right. When P < 0.5, then it is skewed to the left. The skewness of the binomial distribution is more as the gap in P and q widens but becomes less pronounced as ‘n’ increases.

The general form of binomial distribution for ascertaining r’ successes is $P_r = n_c\ r\ P^r q^{n-r}$.

This equation clearly states that in binomial distribution only two parameters are there ‘n’ and ‘P’.

We have explained the binomial distribution only in terms of binomial formula, but the binomial distribution, like any other distribution can be expressed graphically as well.

To illustrate, consider a situation at quality control centre, where a quality control engineer selects five screws at random and experience is 20 percent of the production is defective. To draw the probability distribution of 0, 1, 2, 3, 4 or 5 defective in a sample. We have to apply the binomial formula to calculate these.

P=0.2

q = 0.8

n = 5

For r = 0 i.e. probability of getting a zero defective screw is P (r = 0 = ${}_5C_0\ (0.2)_0\ (0.8)_5$

= 0.3277 for r = 1

5

P(r = 1) = $^5C_0\ (0.2)^1\ (0.8)^4$ ¯ 2 = 0.2 $(0.8)^4$ = 0.4096.

for r = 2

P (r = 2) = ${}_5C_2\ (0.2)_2\ (0.8)_3 = \frac{5\ \ 4}{1 2}\ (0.2)_2\ (0.8)_2 = 0.2048$.

for r = 3

p (r = 3) = ${}_5C_3\ (0.2)_3\ (0.8)_2$

$= \frac{5\ 4}{1\ 2}\ (0.2)_3\ (0.8)_2 = 0.0512$.

for r = 4

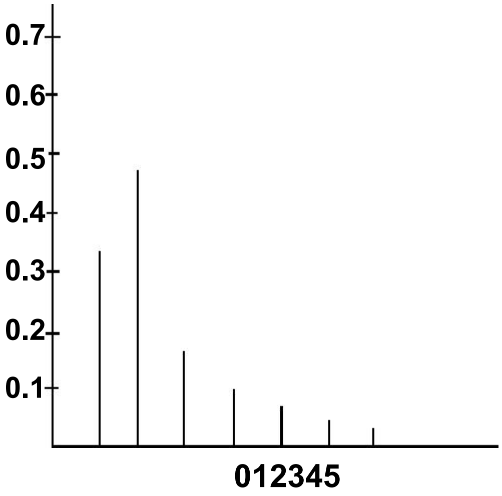
p (r = 4) = $^5C_4\ (0.2)_4\ (0.8)_1$

$\frac{5}{1}\ (0.2)_4\ \ (0.8)_1 = 0.0064$

for r = 5

P (r = 5) = ${}_5C_5\ (0.2)_5 = \text{I.}(0.2)_5 = 0.0003$.

It can be represented graphically as follows :



8.5. Properties of Binomial Distribution :

The binomial distribution has three significant attributes.

They are arithmetic mean, standard deviation and pattern or shape of the distribution.

%4 u The mean of the binomial distribution is Mean = nP

%4 u The standard deviation of the binomial random variable measures the dispersion of the binomial distribution.

The standard deviation is determined by $\sqrt{npq}$

3. Pattern or shape of the binomial distribution.

The shape of the binomial distribution depends upon the value of P and n. If P = q = 0.5 the distribution will be symmetrical regardless the value of n. If P q then the distribution will be asymmetrical. Where P is small (0.1), the distribution is skewed to right as P increases (to 0.3 for example), the skewness is less noticeable When P is larger than 0.5, the distribution is skewed to left. As the value of n increases, the vertical lines not only become more numerous but also tend to bunch up together and form a bell shaped curve.

(A) Self Assessment

State whether the following statements are true or false:

%4 u The study of a population can be done either by constructing an observed (or empirical) frequency distribution, often based on a sample from it, or by using a theoretical distribution.

%4 uIf a random variable satisfies the conditions of a theoretical probability distribution, then this distribution can be fitted to the observed data.

%4 uIt is possible to test a hypothesis about a population, to take decision in the face of uncertainty, to make forecast, etc.

%4 uTheoretical probability distributions can be divided into two broad categories, viz. discrete and continuous probability distributions.

%4 u Binomial distribution is a theoretical probability distribution which was given by James Bernoulli.

%4 uIn Binomial distribution, an experiment consists of a finite number of repeated trials.

<u>EXERCISES</u>

%4 u Over a long period of time, it has been observed that a soldier can hit the target on a single trial with probability equal to 0.8. If he fires four shots at the target find—

(a) probability that he will hit the target exactly two times.

(b) at least two

times. [Hint:

$P(r = 2) = c_2 (0.8)^2 (0.2)^2 = 0.1536$

$P(r < = 2) = P(r = 2) + P (r = 3) + P (r = 4)$

$= {}^4C_2 (0.8)^2 (0.2)^2 + {}^4C_3 (0.8)^2 (0.2)^1 + {}^4C_4 (0.8)^4 = 0.9728$].

%4 u From a incoming production lot, 10 items are examined and the lot is rejected if two or more defective are observed. If a lot contains exactly 5 per cent defectives, what is the$_1$ probability the lot will be

%4 u rejected.

%4 u accepted.,

%4 u Hint : Probability of lot rejected = I-P(r = 0)-P(n=I)

$= 1- {}^{10}C_0 (0.95)^{10} - {}^{10} C_1 (0.05)^1$

$(0.95)^2 = 1- 0.914 = 0.086$]

Given a binomial function.

$$P(r) = {}_{10}C_r \frac{5^{r}}{1} \quad \frac{5_{10-r}}{1}$$

find the median and mode of the distribution.

[Hint: When P = q = $\frac{1}{2}$, distribution is symmetrical and in symmetrical distribution. X = M = Z

Mean = nP = 10 $\frac{1}{2}$ = 5

Give that X is B (n, 0.5), find the mean and variance of X for n = 3, 5 and 10.

[Hint : Mean = nP and variance = nq]

%4 u It is known from the experience that 30 percent of the orders received by a company are placed by XYZ Ltd. If 10 orders are received by a company, find the probability that

%4 u 6 orders are from XYZ Ltd.

%4 u at most two orders from XYZ Ltd.

%4 u at least three orders from XYZ Ltd. [Hint :

%4 u P (r-6) = $^{10}C_6$ $(0.3)^6(0.7)^4$

%4 u P(r2) = P(r = 0) + P (r = 1) + P(r - 2)

%4 u (r $\geq$ 3} = 1 - P (r $\leq$ 2)]

6. Suppose it is known that one out of 10 text book is an outstanding success. A publisher has selected 10 books for publication. What is the probability that

%4 u exactly one will be a outstanding success.

%4 u At least one

%4 u no book get a outstanding success

Hint

%4 u P(r= 1) = $^{10}C_1$ $(0.1)^1$ $(0.9)^9$

%4 u P(r $\geq$ 1) = 1 - P (r = 0)

%4 u P (r = u) - $^{10}C_0$ $(0.1)^0$ $(0.9)^{10}$

%4 u The probability that India wins a cricket match against Pakistan is given to be 1/3. If India and Pakistan play 6 test matches, what is the probability that

(a) India will lose all the six matches.

(b) India will win at least one test match.

(c) India will win at most 2 test matches.

(d) 2 or more than two test matches.

(e) more than 2 test matches.

(0 2 or less than two test matches.

(g) less than two test matches.

(MBA, Delhi 1982; slightly modified)

%4 u 6. Fitting a Binomial Distribution:

The following procedure is adopted while fitting a binomial distribution to the observed data:-

(1) Determine the value of p and q. If one of these values is known the other can be found out by the simple relationship p=1 - q and q = 1- p.

(2) Note the value of n and N. n is the no. of trials in an experiment and N is the total number of trials in all the experiments.

(3) Expand the binomial $(q = p)_n$ to find the probability of all possible number of successes.

(4) Multiply each term of the expanded binomial by N and result will be the required expected frequencies. Illustration:- Four coins were tossed 160 times and the following results were obtained:

No. of heads:	0	1	2	3	4
Frequency:	17	52	54	31	6

Fit a binomial distribution under the assumption that the coins are unbiased.

Solution: Under the assumption that the coins are unbiased, the probability of head (p) and trial (q) are and .

In this case, n = 4, N= 160.

The probability of 0, 1, 2, 3, 4 heads will be given by P(r) = nC_r q_{n-r} p_r.

In order to obtain the expected frequencies, we will have to multiply each probability by N.

The expected frequencies will be obtained as follows:-

Number	Expected Frequency
(n)	$N\ {}_nC_r\ q^{n-r}.\ p^r$
0	$160\ {}^4C_0\ \frac{1}{2}^4\ .\frac{1}{2}^0\ 10$
1	$160\ {}^4C_1\ \frac{1}{2}^3\ \frac{1}{2}^1\ 40$
2	$160\ {}^4C_2\ \frac{1}{2}^2\ \frac{1}{2}^2\ 60$
3	$160\ {}^4C_3\ \frac{1}{2}^1\ \frac{1}{2}^3\ 40$
4	$160\ {}^4C_4\ \frac{1}{2}^0\ \frac{1}{2}^4\ 10$

%4 u Self

Assessment : Fill

in the blanks:

%4 u The purpose of fitting a distribution is to examine whether the observed frequency distribution can be

regarded as a .... from a population with a known probability distribution.

8. To fit a binomial distribution to the given data, we find its ....

9. A distribution is known if the values of its .... are Known.

10. Binomial distribution is often used in various .... Situations in business.

%4 u Acceptance sampling plan, a technique of quality control, is based on ....

%4 u The values of different coefficients, for different values of n, can be obtained directly by using ....

8.7. POISSON DISTRIBUTION

Another important discrete probability distribution is Poisson exponential distribution, named after Simeon Dennis Poisson, a Frenchmen who developed the distribution from studies in 1837. The Poisson distribution is used to describe a number of processes that involves an observation per unit of time or space. For example the number of telephone calls received at a switch board per minutes, number of patients coming to health centre per day, the arrival of cars at a car parking per hour, number of accident at a crossing per day, the number of customers arriving at a bank teller per minute, number of strikes in a factory during one year, number of times a break down in a machinery during three months, floods during the year, accidental release of radiation from a nuclear reactor, number of break downs in a power station per day, number of orders received by a company per day, all follow approximately the Poisson pattern. Such processes are characterized by the expected number of successes per unit of time or space just as the binomial distribution is characterized by the number of successes in 'n' trials.

These examples all have a common element. They can be described by a discrete random variable that takes on integer (i.e. whole number) values 0, 1, 2, 3........so on. The number of break downs in a machine, in a given interval of time will be 0, 1, 2, 3 ....n Number of calls received in a switch of patients arrive in a health centre will be a whole number in a given-time interval.

The above mentioned per unit of time or space is equivalent to the sample size 'n' in a binomial distribution. The binomial distribution is based on ascertaining the probability of a 'r' successes in one trial What, then constitute one trial in a Poisson/distribution. In the case of a batch of parts, an observation of a single part to determine whether it is defective, is of course, a trial. But what about a telephone calls received per minute if we divide the minute into 60 seconds there is still the possibility of receiving more than one call in a given second. In order that this would not happen, if we divide the minute into small sub-divisions such as one hundredths of a second. Then we observe that probability of receiving two or more calls in such a negligible interval of time is so small that it can be ignored for

practical purposes. When unit of time or space is divided into many sub-divisions (n becomes large) and the probability of success on a single trial becomes small. We may take the number of accidents occurring in a given time interval with ‘x’ denoting the average number of accidents per month. For space example, let us take a number of defects per meter of cloth. For each of such situations, the probability of dividing the time or space interval into ‘n’ very small segments such that within a small segment the conditions of Bernoulli process holds, good. As one month can be subdivided into 30 24 60 minutes or 30 24 60 60 second of time intervals each. The probability of occurrence of an accident in this time interval is $\frac{X}{30\ \ 24\ \ 60}$ per minutes $\frac{X}{30\ \ 24\ \ 60\ \ 60}$ or per second. Thus the probability reduces to a small quantity and n becomes very large. Similarly we can subdivide the meter into decimeter, centimeter or millimeter. The main characteristic of Poisson distribution is where n is very large and P is very small and is regarded as a limit of binomial distribution.

The Bernoulli process with a very large number of trial (n) and with a very low probability of success in any trial, then the probability of sample space (0, 1, 2. 3.....n successes) can be ascertained by the formula.

$$P(r) = \frac{e^{-}}{r!}\ .\ ^{r}$$

where e is a constant with a value 2.71828 and the mean value.

Suppose the average number of accidents on a particular crossing to the police record is three accidents per week. The safety division wants to know the probability of exactly 0, 1, 2 and three accidents in a week, e’s negative powers can be ascertained directly form the Poisson distribution table.

Poisson Distribution of Accident per Week

Number of accident		Probability P(r)
0	$e_{-2}\frac{3^0}{01}$	0.0498
1	$e_{-2}\frac{3_2}{1!}$	0.1494
2	$e_{-3}\frac{3^3}{2!}$	0.2242
3	$e_{-3}\frac{3^2}{3!}$	0.2242
4	$e_{-3}\frac{3^4}{4!}$	0.1681

Note for e_{-3} = 0.0498 See statistical tables.

These calculation answer several other questions. Suppose we want to ascertain what the probability of 0, 1, or 2 accidents per week. This can be calculated by adding together the probabilities of exactly 0, 1 and 2 accidents like

P(O) = 0.0498
P (I) = 0.1494
P (2) = 0.2242
P (0, 1, 2) = 0.4234

Suppose we want to ascertain the probability of more than two accidents per week. Then it can be ascertained as follows :

P (r > 2) = 1 - P (O) - P(2) = 1 - 0.4234 = 0.5766

Please note the difference in more than two accident per week [P (r > 2)] and 2 or more than two accidents per week [P (r > 2)] and 2 more than two per week [P (r = 2)]

P(r > 2)= 1 - P(O) - P(1) - P(2)

P(r ≥ 2) 1 - P(O) - P(1)

8.8. Properties of Poisson Distribution:

Like binomial distribution, the Poisson distribution has three significant attributes. They are mean, standard deviation and shape of the distribution.

%4 u The mean of the Poisson distribution is equal to nP (( ). The value of ' ' is usually a small positive number.

%4 u The variance of Poisson distribution is also ' ' The variance of binomial distribution is equal to npq or nP (1-p). As P approaches O, the limit of the last factor (1 - P) approaches 1 and variance approaches nP or . The standard deviation of Poisson distribution is equal to $\sqrt{\ }$ or $\sqrt{nP}$. From this we can draw a inference that Poisson distribution has only one parameter i.e. .

3. Shape of the Poisson Distribution:

The Poisson distribution is positively skewed (P being small). Given the value of P, nP, will increase with increase in n. As or nP increases, the Poisson distribution will be closer and closer to bell shaped distribution. That is why, normal distribution is also the limit of the Poisson distribution.

%4 u Self Assessment: Fill in the blanks:

The purpose of fitting a distribution is to examine whether the observed frequency distribution can be regarded as a .... from a population with a known probability distribution.

8. To fit a binomial distribution to the given data, we find its ....

9. A distribution is known if the values of its .... are Known.

10. Binomial distribution is often used in various .... Situations in business.

11. Acceptance sampling plan, a technique of quality control, is based on ....

12. The values of different coefficients, for different values of n, can be obtained directly by using ....

EXERCISES

1. The average price hike in a car is 4 times in every three years, find the probability of

(a) No price hike in three years

(b) two price hike in three years

(c) 4 price hike in three years

[Hints :

(a) P (r = 0) = $e^{-4}\frac{4_0}{0!}$

(b) P (r = 2) = $e^{-4}\frac{4_2}{2!}$

(c) P (r = 4) = $e^{-4}\frac{4_4}{4!}$

%4 u In a hospital with 20 X-ray machines and the chances of mal-functioning during a day is 0.02. What is the probability that three machine in a day will be out of order.

[Hint :Mean = nP = 20 0.02 = 0.4

$$p(r = 3) = e_{0.4}\frac{0.4^{3}}{31} = 0.00715]$$

3. The one percent of the holes done by a drilling machine are defective. A random sample of 300 holes are examined. Find the probability that

(a) All holes are good

%4 u two or fewer holes are defective

%4 u more than two are defective

[Hint : Mean = nP = 300 x 1/100 = 3

$$\text{(a) } P(r = 0) = e^{-3}\frac{3_0}{0!}$$

(b) P (r = 2) = P (r = 0) + P (r = 1) + P (r = 2)

$$= e_{-3}\frac{3_0}{0!} \quad \frac{}{1!} \quad e_{-3}\frac{3_1}{2!} \quad e_{-3}\;3_2$$

(c) P(r = 2) = 1 - P (r = 0) - P (r = 1) -P (r = 2)]

%4 u In a college 10 per cent of the students are girls. A sample of 50 students are selected at random. With the help of Poisson distribution. Find the probability of that a sample contains.

%4 u all boys

%4 u all girls

%4 u one girl

%4 u less than three girls

%4 u more than three girls

%4 u In a university with 300 teachers, the average proportion of teachers absent in a say is 0.04 per cent. Find the probability that in a given day.

%4 u all teachers are present

%4 u two are absent

%4 u three or more teachers are absent

%4 u three or fewer

teachers are absent Mean = nP =
300 0.04 = 12

$$P(r = 0) = e^{-12}\frac{12_0}{0!} = 0.00001$$

$$P(r = 2) = e^{-12}\frac{12_2}{2!} = 0.00001\frac{144}{2}$$

8.9. Poisson Distribution as Approximation of Binomial Distribution:

Sometimes, the calculation of binomial distribution are very tedious because ‘n’ is very large and P is very small i.e. when number of trials are large and probability of success is very small, Poisson distribution can be used as approximation of binomial distribution. To illustrate, an insurance company hold a large number of life insurance policy on individuals of any particular age, and the probability that a single policy will result in a claim during the year is very low. The distribution of the number of claims is binomial with large ‘n’ and very small P.A company may have

a large number of machines working on a process simultaneously. If the probability that any one of them will break down in a single day is small, the distribution of the number of daily break down is binomial with large n and small P. In such cases, binomial distribution can be approximated by the Poisson distribution. Illustration A large Plant has 50 similar machines operating simultaneously. If the probability that any one of them breaks down in a day is 0.04. Find the probability that 2 machines breaks down

n = 50

Probability of break down - 0.04

%4 u $(r) = {}^{n}C_{r}\, p^{r}\, q^{n-r}$

%4 u $(r = 2) = {}^{50}C_{2}\, (0.04)^{2}\, (0.96)^{48}$

It is a tedious calculations, hence it can be solved with the help of Poisson

distribution which is $P\ (r) = e^{-np}\ \frac{(nP)}{r!}$

nP = 50 0.04 = 2

$P(r = 2) = \frac{e^{-2}\ 2^{2}}{2!} = 0.2706$

Similarly for 2 or more break downs, under binomial distribution will be

$P(r \geq 2) = P\ (r = 2) + P\ (r = 3) + P\ (r = 4) + ...+ p\ (r = 50)$

$+ p\ (r = 50)$

%4 u $1 - P\ (r = 0) - P\ (r = 1)$

%4 u $1 - {}^{50}C_{0}\ (0.04)^{0}\ (0.96)^{50} – {}^{50}C_{1}.(0.04)^{1}\ (0.96)^{49}$

which is again a tedious job, with the help of Poisson distribution, the calculations becomes more easy. $P(r \geq 2) = 1 - P(r = 0) - P(r= 1)$

$= 1-1 - \frac{e^{-2}\ 2^{0}}{01}\ \frac{e^{2}\ 2^{1}}{1!} = 1 - 0.1353 - 0.2706 = 0.5941.$

The rule of thumbs is, Poisson is a good approximation of the binomial distribution when n is equal to or greater than 20 and P is equal to or less than 0.05. In these conditions, we can substitute the mean of the binomial distribution (nP) in place of the mean of the poisson distribution ( ).

(C)Self Assessment

State whether the following statements are true or false: d

%4 u Poisson distribution was derived by a noted mathematician, Simon D. Poisson, in 1837.

%4 u Poisson distribution is a limiting case of binomial distribution, when the number of trials n tends to become very large and the probability of success in a trial p tends to become very smail such that their product np remains a constant.

19. Poisson distribution is used as a model to describe the probability distribution of a random variable defined over a unit of time, length or space.

%4 u The expected number of occurrences in an interval is constant.

%4 u It is possible to identify a small interval so that the occurrence of more than one event, in any interval of this size, becomes extremely unlikely.

EXERCISES

%4 u Between the hours 4 PM and 5 PM, the average number of calls per minute coming to the switch board of a company is 3. Find the probability that during one particular minute there will be no phone call at all.

%4 u Certain mass produced articles of which 0.5% are defective, are packed in cartons each containing 130 articles. What proportions of cartons are free from defective articles? What proportions of cartons contain 2 or more defective?

[Ans : a. P = 0.61, P = 0.09].

%4 u Find the probability that at most 5 defective bolts will be found in a box of 200 bolts if it is known that 2 percent of such bolts are expected to be defective. (Assume Poisson distribution)

[Ans : P = 0.784].

%4 u A manufacture who produces medicine bottles, find that 0.1% of the bottles ae defective. The bottles are packed in boxes containing 500 bottles. A drug manufacture buys 100 boxes from the producer of bottles. Using poisson distribution and how many boxes will contain.

%4 u no defective

%4 u at least two defectives (Ans : a. 61% ; b. 9%)

%4 u If the probability that an individual suffers reaction from a given serum is 0.001. Determine the probability that out of 2000 individuals.

%4 u exactly 3

%4 u more than 2 individuals differ from reaction. [Ans. A. 0.18 : b. 0.323]

(MBA 1987 Jodhpur)

8.10. Fitting a Poisson Distribution:

The following procedure is adopted while fitting a Poisson distribution to the observed date.

%4 u Calculate the value of i.e. average occurrence.

%4 u Calculate the frequency of 0 success.

%4 u Other expected frequencies can be calculated as follows:-N (P_0) = N.e.

N (P_1) = N (P_0) 1

N (P_2) = N (P_1) 2

N (P_3) = N (P_2) 3 etc

%4 u Self Assessment Fill in the blanks:

18. Poisson distribution is .... probability distribution.
19. Poisson distribution has ....
20. The Poisson distribution is a .... distribution.
21. Poisson distribution is applicable to situations where the number of trials is .... and the probability of a success in a trial is ....

8.11. Summary:

Theoretical distribution play an important role in statistical theory. These are useful in analyzing the nature of given distribution under certain assumption. They help us in making logical decisions, making predictions, projection and forecasting theoretical distributions are useful in solving many business and other problems.

The binomial probability distribution is a discrete probability distribution that is useful in describing an enormous variety of real life event. The poisson distribution is used in a wide variety of problems. It is used in quanlity control statistics to count the number of defects of an item, in biology to count the number of bacteria, in insurance problems, to count the number of typing errors per page., etc.

8.12 Glossary

Binomial distribution: Binomial distribution is a theoretical probability distribution which was given by James Bernoulli.

Experiment: An experiment consists of a finite number of repeated trials Fitting of a binomial distribution: The fitting of a distribution to given data implies the determination of expected (or theoretical) frequencies for different values of the random variable on the basis of this data.

Posteriori inferences: These are the basis of results.

Priori considerations: These are the basis of given conditions.

Theoretical probability distribution: A theoretical probability distribution gives us a law according to which different values of the random variable are distributed with specified probabilities.

Poisson Approximation to Binomial: Poisson distribution can be used as an approximation to binomial with parameter m = np.

Poisson Distribution: This is a a limiting case of binomial distribution, when the number of trials n tends to become very large and the probability of success in a trial p tends to become very small such that their product np remains a constant. This distribution is used as a model to describe the probability distribution of a random variable defined over a unit of time, length or space.

Probability Mass Function: The probability mass function (p.m.f.) of Poisson distribution can be derived as a limit of p.m.f. of binomial distribution when n such that m (= np) remains constant.

8.13 Answers: Self Assessment

1. True 7	2. True	3. True
4. True	5. True	6. True
7. sample	8.mean	9. parameters
10. Decision-making	11. Binomial distribution	12. Pascal Triangle
13. True	14. True	15. True
16. True	17. True	18. discrete
19. only one parameter m	20. positively skewed	21. large, very small
22. Refer to section 8.5	23. Refer to section 8.8	

8.14 Terminal Questions

%4 u What do you understand by theoretical frequency distribution?

%4 u What is Binomial distribution? What are its important properties?

%4 u What is Poisson Distribution? Explain the properties of Poisson distribution.

%4 u Determine the binomial distribution (i.e. n, p and q) whose mean is 10 and variance is 8.

%4 u Four coins were tossed 200 times. The number of tosses showing 0,1,2,3 and 4 heads were observed as under:

No. of heads:	0	1	2	3	4
No. of tosses:	15	35	90	40	20

Fit a Binomial Distribution to these observed results.

6. The following mistakes per page were observed in a book:

No. of mistakes per page:	0	1	2	3	4
No. of page:	211	90	19	5	0

Fit a Poisson distribution for the data.

8.15. Further Readings:

%4 u Chao. Lincoln I..- Statistical Methods and Analysis, Me r Yav, Hill.

%4 u Gangolli, R.A. and D. Yulviskakr, Discrete Probability Harcourt, Brace & World Ince New York.

%4 u Levid. R.I.. Statistics for Management, Pretice Hall Inc. New York.

%4 u Mendenhail and Reinmuth, Statistics for Management and Economics Duxbury

%4 u New bold, Paul, Statistics tor Business and Economics, Prentice Hall inc., Englewood Cliffs. New Jersey.

*****

Lesson-9
Normal Distribution

Structure:

9.1. Learning Objectives: After studying this lesson, you should be able to understand.

%4 u Meaning of Normal Distribution
%4 u Significance of Normal Distribution
%4 u Calculation of Area under Normal curve
%4 u Properties of Normal Distribution
%4 u Methods of Calculating Expected frequencies.
%4 u Introduction:

In the previous chapter we have discussed the two important discrete probability distributions i.e. binomial and Poisson distribution that have a finite or countable infinite number of points in a sample space. In this chapter we will discuss the random variable that can take any value on a continum. It has an infinite or uncountable infinite number of values on its sample space. Such random variables are called continuous random variables and their distributions are called continuous distributions. Many types of data are continuous in nature because the observations are obtained by measurements such as time, distance, or temperature, height, weight etc In continuous variables, the successive observations may differ by infinitesimal amounts. The weight of a person, for instance, may be 62 Kg, 62.2 Kg, 62.253 Kg etc. depending upon the accuracy of the weighing machine. Even the most precise instrument available can give measurements of only a limited degree of precision. Thus in actual practice we usually can never have a distribution such that successive observations have gaps in between. Consequently in the measurement of time, distance, height, weight variables, the values obtained are really rounded numbers or approximation. Each of these values represents not a point but an interval in which the measurement falls.

In short, any random variable whose values are measured as a continuous variable whereas in discrete variable it is counted. The probabilistic model of a continuous variable involves the selection of a curve usually smooth called the probability distribution or probability density function. The distributions may assume a variety of shapes. It is interesting to note that very large number of random variables observed in nature possess a curve which is approximately bell shaped and is commonly referred as normal probability distribution. This occupies a central position in statistical inference theory and is widely applied in practical situations.

9.3. Significance of Normal Distribution:

1. Large Number of phenomenons tends to follow a normal distribution pattern.

Numerous variables tends to follow a pattern of variation that is similar to the normal distribution. Natural trials such as heights, weights, I.Q. of children, test scores in a exam, life of electric lamp, breaking strength of a steel rope, errors of measurement, life of an automobile tyre, etc. can be approximated through normal distribution.

%4 u Sampling distribution of many sample statistics such as mean has an approximately normal distribution. Many distributions in business, economic, sociology and other social sciences, do not resemble the normal distribution but the distribution of sample means (sampling distribution of mean) in each case, usually is normal as long as the sample size is large. This property has a wide applications in sampling estimation and testing of hypothesis.

%4 u Other types of sampling distributions can be approximated by normal function. In the previous chapter, we observed that binomial and Poisson distribution approach to normal distribution when the value of ‘n’ is very large. It is too laborious to work out the probabilities for a binomial random variable or Poisson variable when ‘n’ is very large. In these conditions, normal distribution can be used as approximation to binomial distribution.

For these reasons, normal distribution is indispensable and is widely used in practice.

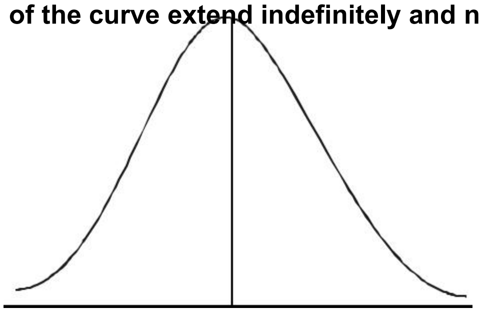
9.4. Normal Probability Distribution :

The normal distribution was first discovered by Abraham De Moivre in 1733. He derived the distribution on the limiting form of the binomial but his work had not been discovered when the same formula was derived by Karl Friedrich Gauss. He first made reference to it in 1809. He was a great mathematician cum astronomer. In honour of his work, the normal probability distribution is often called the Gaussian distribution. He applied the normal function in evaluating errors in observations in astronomy. Thus the normal probability distribution is often referred as normal function of error.

9.4 (a) Normal Curve:

The shape of the normal curve is bell shaped as stated earlier graphically it is.

The curve has a single peak thus it is unimodel. The mean of a normally distributed population lies at the centre of its curve. Because of the symmetry, the median (positional average) and mode (concentration of frequency) also lies at the centre. Thus in normal distribution mean, median and mode all coincide. The two tails of the curve extend indefinitely and never touches the X-axis.



—Mean

—Median

—Mode

9.4 (b) Normal Functions :

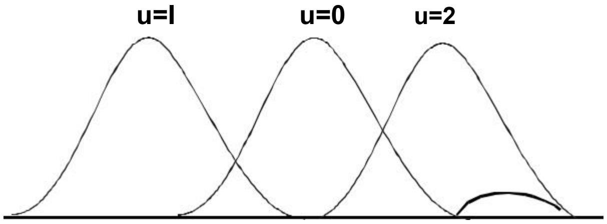
The binomial probability function is $P(r) = {}_{n}C_{r}\ P_{r}\ q_{n-r}$ derived from $(q + P)_{n}$. When ‘n’ becomes infinitely large, this function becomes normal function in the following form.

$$P(x) = \frac{1}{2\pi.} .e^{-\frac{1}{2}\frac{x^2}{2}}$$

Where e = 2.71828 and = 3.1416 are constants. The P(x) stands for the probability density at each possible value of x and is the ordinate of the curve at each possible point x on axis. The entire area under the curve is equal to I. The above equation clearly indicates that normal function has only two parameters i.e. mean ( ) and variance ( ). When the parameter ‘n’ of binomial distribution is very large then its mean (nP) will be equal to ‘ ’ and variance (nPq) is equal to $_{2}$.

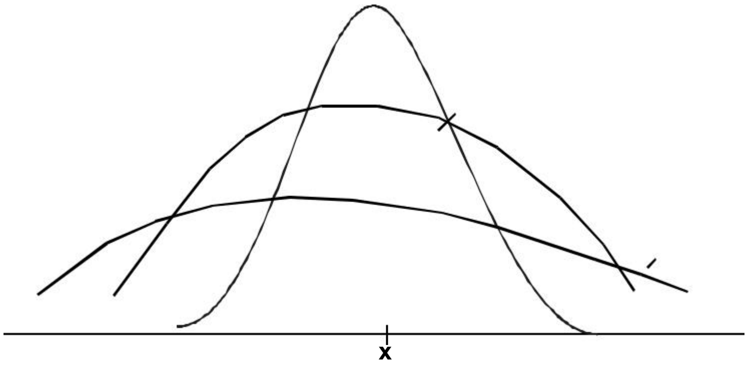
9.4.(c) Parameters :

As we noticed, there are only two parameters of normal distribution i.e. mean ( ) and variance ($_{2}$). The mean determines the location of the curve and variance determines the shape of the curve. Given variance, a change in mean will shift the curve as a whole along the x-axis. The following figure depicts the three normal curves, having same variance but different means.



On the other hand, given mean, a change in variance will change the shape of the curve. The following figure depicts the three normal curves having same mean but different variance. Higher the variance reveals higher dispersion or scatterness around the central value i.e. mean and vice versa. If all the observations are same then variance will be zero and all lies on the same point i.e. mean. Higher the dispersion, the flatter the curve will be, the smaller the dispersion, the peaked the curve will be.

That's why, the normal distribution is commonly denoted by N ($_{2}$). Upto this stage we have explained to the readers normal curve function and parameters. Now we will come to the area under normal curve.



9.5. Area under Normal Curve:

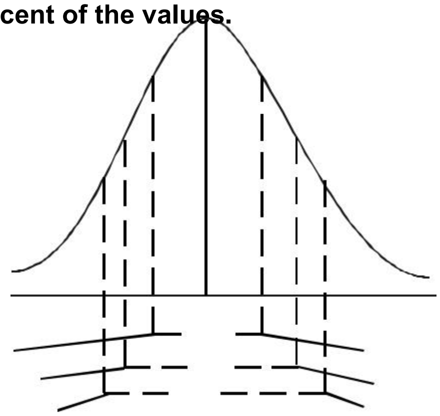
As explained earlier, the total area under the curve is equal to 1.00 regardless the values of mean and variance. The area under the curve can be treated as probabilities as total area is one. Because, the sum of probabilities of a sample space is one. Mathematically it is true that

%4 u Covers 68.27 percent of the values in a normally distribution population.

%4 u Covers 95.45 percent of the values.

%4 u Covers 99.73 percent of the values.

Graphically



In practice, very few of the applications we shall make of the normal distribution involve intervals of exactly, 1, 2 or 3 standard deviation form the mean. What should we do when we wanted to ascertain the area under normal curve involving or 0.7 In these situations, we use normal distribution tables for ascertaining the probabilities. It is not possible to have a different tables for every possible value of mean and variance. Instead, we convert the normal distribution into a standard normal probability distribution. Table for standard normal probability distribution are available to find areas or probabilities under any normal curve.

9.6. Standard Normal Distribution :

In this sub-section, we will study how a particular normal distribution is transformed into a standard normal distribution. This special distribution has a mean 0 and a standard deviation I and is written as N (0, 1).

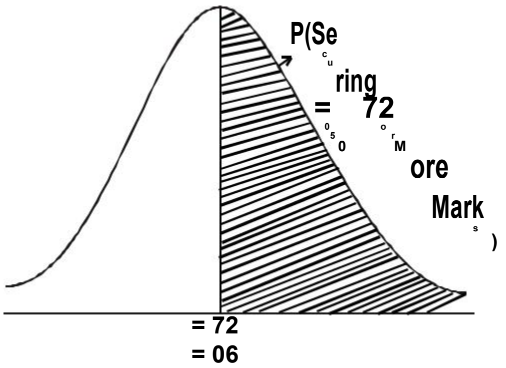
Standard normal distribution is the distribution of another normal variable called Z-scores, which is defined as, $z = \frac{\overset{-}{x}}{\quad}$.

The Z-score is the difference of an observation X from the mean ( ) expressed in term of standard deviation ( ). It is called Z-scores because random variable take on many different units of measure e.g. inches, rupees, degrees, hours, etc. Since we use only one table which should be standard unit and is represented by Z this table gives the value or area of probability between variable and mean.

Illustration:

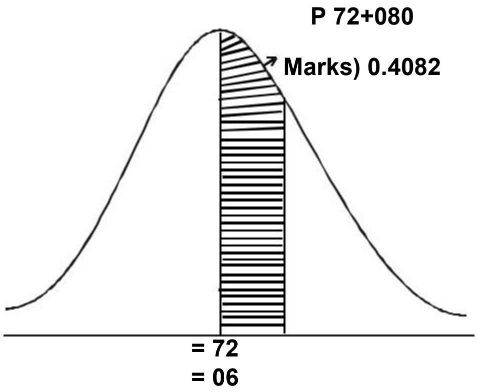
Suppose a standard test gives a mean score 72 with a standard deviation 6.

Class I — What is the probability that a student selected at random will score 72 or more marks.



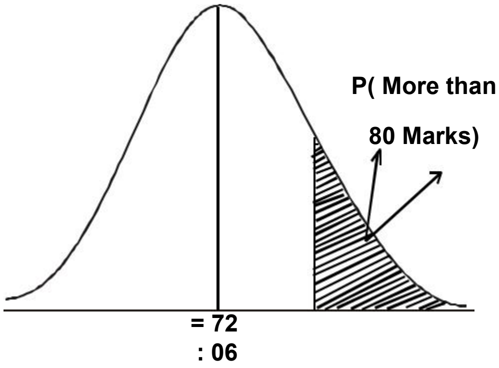
We know that half of the area under the curve is located on either side of the mean 72. Thus, we can deduce that a student selected at random will score 72 or more marks is the shaded area i.e. 0.5.

Case II—Suppose we wanted to ascertain the probability that a student will score between 72 and 80 marks. Using Z-score equation which is



If we go through the standard normal distribution table, the area of probability corresponding to 1.33 is 0.4082.

Case III—What is the probability, that a student will score more than 80 marks.



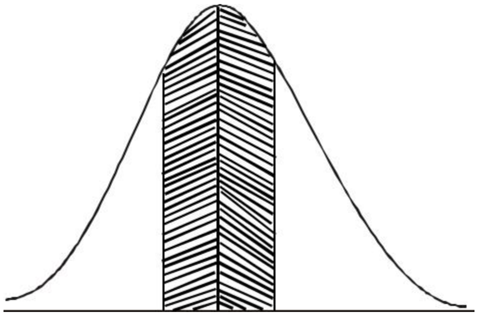
The standard normal tables gives the area under the curve between variable and mean. Now we wanted to ascertain the value of the shaded area. We know the area to the right side of mean is 0.5 and area between 72 to 80 marks is 0.4082 (as calculated above). Hence the shaded area will be

Area to the right of mean = 0.5000
Area between 72 to 80 = 0.4082
Area to the left of 80 = 0.0918 (i.e. 0.5000—0.4082)

The probability that a Candidate will score more than 80 marks is 0.0918.

Case IV — Now, suppose we wanted to ascertain the probability, that the marks obtained will be between 63 and 79.

In this case, firstly, we will calculate the area between 63 and 72. Then area between 72 and 79. Then add both the area.



$$Z_1 = \frac{63\quad 72}{6}\quad \frac{9}{6}\ 1.5$$

The Z—value of 1.5 has a probability of 0.4332.

The Z — value of 1.16 has a probability of 0.3770.

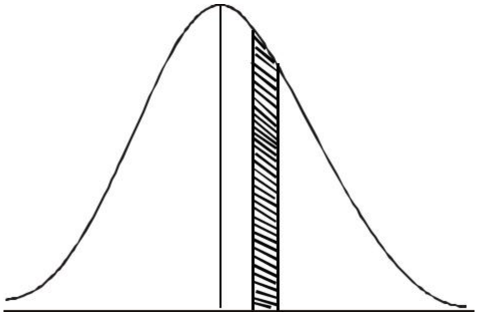
The probability that marks will be between 63 and 72 = 0.4332.

The probability that marks will be between 72 and 79 = 0.3770

The probability that marks will be between 63 and 79 = 0.8102

Case V — Suppose we want to estimate the probability of marks obtained between 75 and 85. Now we are interested in the shaded areas as depicted in the following figure.

In this case, firstly we will determine the area of probability between 72 and 85 i.e. a, area between 72 and 75 i.e. b, the area between 75 and 85 will be equal to (a-b).



Area between 72 and 75 is $Z_1 = \frac{7572}{6} \quad \frac{3}{6} \, 0.5$.

The probability Corresponding to Z score of 0.5 is 0.1915.

Area between 72 and 85 is $Z_1 = \frac{8572}{6} \quad \frac{13}{6} \, 2.16$.

The probability corresponding to Z-score of 2.16 is 0.4846.

The probability of marks obtained between 75 and 85 will be 0.4846 - 0.1915 = 0.2931.

9.7. Properties of Normal Distribution Curve :

Normal curve is unimodel and symmetrical. Because of this, it has only one largest ordinate and mean=median=mode.

The two tails of the normal distribution extend indefinitely and never touches X-axis and limits areto +

The first and third quartile are equidistant from the median i.e. Q_3 M = M Q_1.

A random variable, discrete or continuous, skewed or symmetrical, the sampling distribution of mean will have approximately normal distribution if the sample size is sufficiently large. This most remarkable and immensely significant fact is known as ‘Central Limit Theorem.’

(A)Self Assessment

State whether the following statements are true or false:

%4 u The Normal distribution was first observed as the normal law of errors by the statisticians of the eighteenth century.

%4 u The observed distribution of a random variable was found to be in close conformity with a continuous curve, which was termed as the normal curve of errors or simply the normal curve.

%4 u p and e are absolute constants with values 3.14159.... and 2.71828.... respectively.

%4 u Normal probability distribution is a bell shaped symmetrical curve about the ordinate at X.

%4 u Normal probability distribution is unimodal curve and its tails extend ‘infinitely in both directions.

%4 u In Normal distribution, since the distribution is symmetrical, all odd ordered central moments are zero.

EXERCISES:

The length of the life a water tap washer is approximately normally distributed with mean 3.5 year and standard deviation 1.1 year. If this type of washer is guaranteed for one year what fraction of original washers sold will require replacement ?

[Hint : Find the probability that a washer will last less than one year. Only these washers will come for replacement.

$$Z = \frac{1 \quad 3.5}{1.1} \quad 2.27.$$

The area with respect to 2.27 is equal to 0.4884. Thus the area to the left of 1 year will be 0.5000-0.4884=0.0116.

1.16% of the sales will come for replacement].

The life of a tyre is normally distributed with mean 38000 miles and standard deviation 5000 miles. For a randomly selected tyre, find the probability that its life will be

%4 u less than 40,000 miles

%4 u more than 42,000 miles

%4 u less than 32,000 miles

%4 u more than 30,000 miles

%4 u between 32,000 miles to 42.000 miles

%4 u between 25,000 miles to 36,000 miles

%4 u The average life of a bulb is 2000 hrs with a standard deviation of 400 hours. One bulb is randomly selected from the production lot. Find the probability that the life of the selected bulb will be

(a) at least 1000 hours.

(b) at most 3000 hours.

(c) between 1500 to 2800 hours.

%4 u The breaking strength of a plastic bag is normally distributed with mean 5 kg and standard deviation 0.8 kg. What proportion of the bags will break at 6.8 kg.

%4 u Suppose that distribution of weekly wages of 10,000 production workers is normal and has a mean of 110 and variance 64. How many workers have wages that are

(a) equal to or less than 100

(b) equal to or less than 125

(c) equal to or more than 90

(d) equal to or more than 120.

%4 u 10,000 light bulbs with a mean life of 180 days are installed in a locality with a standard deviation of 30 days. How many bulbs will expire in less than 100 days.

%4 u The diameter of a metal rod is normally distributed with mean 4cm and standard deviation 0.1cm.

%4 u what proportion of the rods have diameter more than 4.16.

%4 u what proportion of the rods will have diameter less than 3.80cm.

9.8. Normal Distribution as an Approximation to Binomial Distribution :

The normal distribution is a continuous distribution but in certain cases it can be used to approximate discrete distributions. When sample size 'n' increases, the binomial probability distribution approaches the smooth bell-shaped form. If 'X' a random discrete variable is approximately normal, its value can be transformed into a Z-score as follows:

$$Z = \frac{X-E(X)}{\sqrt{nPq}} \quad X-nP$$

Precaution should be taken when transferring a discrete binomial variable into a continuous standard normal variety, a correction for continuity be applied. The binomial variable assumes the values in positive integers such as 0, 1, 2, 3....... so on whereas the probabilities of a continuous variable random variable are stated in terms of intervals. Consequently, a correction factor is applied.

Case 1 The probability of exactly 'Y' successes is approximated by evaluating the area under normal curve between Z_1 and Z_3.

$$Z_1 \quad \frac{(Y-0.5)-(nP)}{\sqrt{nPq}}$$

and $$Z_2 \quad \frac{(Y_2 +0.5) \quad (np)}{\sqrt{nPq}}$$

Case II The probabilities of between and including Y_1 and Y_2 successes is approximated by computing the area under normal curve by applying correction for continuity. In lower limit we subtract 0.5 and in upper limit we add 0.5.

$$Z_1 \quad \frac{(Y-0.5)-(nP)}{\sqrt{nPq}}$$

and $$Z_2 \quad \frac{(Y_2 +0.5)-(nP)}{\sqrt{nPq}}$$

Case III Similarly in case of more than or less than we make correction of 0.5 for continuity.

Illustrations :

The production line of a large manufacturing plant produces items of which 10 percent are defective. In a random sample of 100 item selected at random from the production line, what is the probability of getting.

%4 u exactly 5 items defective.

%4 u between 8 to 12 defective items.

%4 u 12 or more than 12 defective items.

%4 u 8 or less than 8

defective items. [Hint

%4 u The problem can be solved with the help of binomial distribution but the calculations will be very lengthily for example

n = 100 P = 010 q = 0.9

P(r=5) = ${}_{100}C_{5}\,(0.1)_5\ (0.9)_{95}$ = 0.0338658

If we solve with the help of normal approximation method then mean = nP = 100 0.10 = 10

variance = nPq = 100 .1 .9 = 9

standard deviation = 3

The probability of exactly 5 would be approximated by the area under the normal curve between 4.5 and 5.5

$$Z_1 \quad \frac{4.5\text{-}10}{3} \quad \frac{5.5}{3}\ 1.83$$

$$Z_1 \quad \frac{5.5\text{-}10}{3} \quad \frac{4.5}{3}\ 1.5.$$

The area between 4.5 and mean is = 0.4664.

The area between 5.5 and mean is = 0.4332

The area between 4.5 and 5.5. is = 0.0332

(b) between 8 and 12.

The binomial distribution provides

P (8 r 12) = P (8) + P (9) + P (10) + P (11) + P (12)

which is again a tedious job, the normal distribution provides

$$Z_1 \quad \frac{(8\text{-}0.5)\text{-}10}{3} \quad \frac{7.5\ \ 10}{3}\ 0.83$$

and

$$Z_2 \quad \frac{(12+0.5)\text{-}10}{3} \quad \frac{12.5\ \ 10}{3}\ 0.83$$

The area with respect to 0.83 is 0.2967.

Hence the area between 8 and 12 is 0.2967 + 0.2967 = 0.5934.

(c) 12 or more than 12 defective items

The probibility of P (r $\geq$ 12) = P 912) +P (13) + (14) + ...+ P (99) + P (100).

The normal approximation to P provides find the are under the normal curve to the right of (12-0.5) = 11.5.

$$Z_1 \quad \frac{11.5=10}{3} \quad \frac{1.5}{3}\ 0.50$$

The area corresponding to Z-score of 0.50 is 0.1915. Hence, the area of probability of more than 11.5 is 0.5000-0.1915=0.3085.

(d) 8 or less than 8 defective items

The probability is P (r $\leq$ 8) = P(0) + P(1) P(2) + ... P(8)

The normal approximation is $Z_1 \frac{}{3} \frac{7.5\text{-}10 \quad 2.5}{3} \quad 0.83$

The area between mean and 7.5 is 0.2967. The area to the left of 7.5 is equal to 0.5000-00.2967 = 0.2033. EXERCISES

%4 u A random sample of 1000 fuses was tested and 2 percent were found defective. Calculate the probability of

(a) exactly 30 were defective

(b) more than 30 were defective

(c) less than 10 wrer defective

%4 u A manufacturing process which produces election tubes is known to have a 5% defective rate. if a sample of 25 is selected from a manufacturing process. Find the probability that

(a) two tubes are defective

(b) between 1 and 2 tubes are defective.

%4 u On the basis of past experience, automobile inspector of a service station noticed that 5 per cent of all cars coming to the station for inspection fail to pass. Using the normal approximation to the binomial distribution. Find the probability.

(a) that between 5 to 15 of the next 200 cars coming to the station fail in the inspection.

(b) less than 9 cars fail in the inspection.

(c) more than 12 cars fail in the inspection.

%4 u Fitting a Normal Curve

There are two objects of fitting a normal curve to sample data.

%4 u To provide a method for judging whether or not the normal curve is a good fit to sample data.

%4 u To use the smoothed normal curve for estimation i.e. to study the characteristics of population.

%4 u Methods of Fitting: There are two methods of fitting normal curve

(1) Method of Ordinates

(2) Method of Areas

1). Method of Ordinates:-

White fitting a normal curve the ordinates are obtained at various sigma distances from the mean. The procedure of obtaining ordinates is a follows:-

The height of the mean ordinate $y \quad \frac{Ni}{2.5066} \, 2.71828^{\frac{1}{} \left(\frac{X - \overline{X}}{}\right)^2}$

We need the values of N, $\overline{X}$ and in order to fit a normal curve to a distribution. When (X - $\overline{X}$) = 0, the exponent

of 2.71828 raised to the zero power is one thus the expression $e^{x_2 / 2 \, 2}$ is always equal to 1 for the ordinate

erected at the mean. The mean ordinate can be now $y_0 \quad \frac{Ni}{2.5066}$ or $y_0 \quad 0.399 \frac{Ni}{}$.

This is the maximum ordinate of the fitted curve. The height of the ordinate at a distance 1 from the mean would

be calculated as follows: $y_1 \quad 0.399 \frac{Ni}{} e^{\frac{1}{2}(1)^2}$.

In a similar manner the height of the ordinate at a distance of 2 from mean would be calculated. Heights of the ordinates can be found out from a specially prepared mathematical table.

%4 u Method of Area: The area under the normal curve represents the total number of frequencies. When the method of areas is used to fit the normal curve we obtain areas (frequencies) with the various class intervals. Tables would be consulted which give area under the normal curve.

Illustration: Fit a normal curve to the following data by (a) Method of ordinates (b) Method of areas

Variable	Frequency
60 62	5
63 65	18
66 68	42
69 71	27
72 74	8
	N= 100

Solution: Since the given series is inclusive in form, it must be converted into exclusive series and real class intervals will be obtained as:

Class	Frequency	Mid Point	$\frac{X-67}{3}$		
	F	X	d	fd	fd_2
59.5 62.5	5	61	2	10	20
62.5 65.5	18	64	1	18	18
65.5 68.5	42	67	0	0	0
68.5 71.5	27	70	+1	27	27
71.5 74.5	8	73	+2	16	32
	N=100		fd=15		fd_2 =97

$$\bar{X} = A + \frac{fd}{N} i \quad 67 \quad \frac{15}{100} 3 \;\; 67.45.$$

$$\sqrt{\frac{fd^2}{N} \quad \frac{fd}{N}^{2}} i \quad \sqrt{\frac{97}{100} \quad \frac{15}{100}^{2}} \; {}^{3}.97 \;\; (.15)^2 \sqrt{\qquad\qquad} \quad 3 \;\; 2.92.$$

Therefore, the value of mean ordinate, i.e. Y_0 can be obtained as:

$$Y_0 = 0.399 \frac{100 \;\; 3}{2.92} = 4.99 = 41.$$

Method of Ordinates:

Now, the following table will show the height of ordinates obtained at the midpoint of different class intervals

Classes	Frequency	Mid Point	X- $\bar{X}$ X-67.45	$\frac{x}{}$	Proportionate Ordinate	Expected Frequency
f	X				$\frac{1}{2} \frac{x}{}^{2}$	
59.5 62.5	5	61	6.45	2.209	0.0869	356 or 4
62.5 65.5	18	64	3.45	1.182	4.989	20.40 or 20
65.5 68.5	42	67	0.45	0.154	9.881	40.52 or 41
68.5 71.5	27	70	+2.55	+0.873	6.849	28.01 or 28
71.5 74.5	8	73	+5.55	+1.901	1.644	6.74 or 7
	N=100					

The following procedure is adopted:

%4 u Take the midpoint of different class intervals.

%4 u ¯Find X - X i.e. x.

%4 u Divide (X - X) or by (S.D.)

4. With the help of X
X find the ordinates from the table.

———point ‘ ’ with the mean ordinates. It will provide expected frequencies.

%4 u Multiply the ordinates at Method of Areas:

Class intervel	Frequencies f	Lower Class limit (x)	X-X or x-67.45x	x —	Area under Normal from mean 0 to Z	Area for end class	Expected Frequencies
1	2	3	4	5	6	7	8
59.5-62.5	5	59.5	-7.95	-2.72	0.4967	0.0413	4.13 or 4
62.5-65.5	18	62.5	-4.95	-1.70	0.4554	0.2068	20.68 or 21
65.5-68.5	42	65.5	-1.95	-0.67	0.2486	0.3892	38.92 or 39
68.5-71.5	27	68.5	+1.05	+0.36	0.1406	0.2771	27.71 or 28
71.5-74.5	8	71.5	+4.05	+1.39	0.4177	0.0743	7.43 or 7
		74.5	+7.05	+2.41	0.4920		
	N=100						

The following procedure is adopted:

%4 u Give the real limits of class intervals

%4 u Take the lower limits of the different groups.

%4 u Find X - X

x

%4 u

%4 u Find the area for different class intervals from table.

%4 u Column 7 is a column of difference. It gives areas under normal curve between successive values of Z.

%4 u Find expected frequencies by multiplying the area of column ‘7’ with total of the frequencies, i.e. 100.

(B)Self Assessment Fill in the blanks: :

%4 u Normal distribution can be used as an approximation to binomial distribution when n is large and

.... p .... q is Very small.

8. In Normal distribution, the standard normal variate z would vary from .... to ....

9. A .... is fitted to estimate the characteristics of the population.

10. The fitting of a normal curve can be done by the..

11. Under Method of Areas, the .... or the areas of the random variable lying in various intervals are determined.

%4 uSelf Check Exercise:

What do you mean by Normal Probability Distribution?

What are the parameters of normal distribution?

What are the properties of Normal Distribution Curve?

9.11. Summary:

Normal distribution has occupied a very important role in Statistics. It is generally used in statistical Quality control in industry in setting up control limits. For large values of n, compilation of probability for discrete distributions becomes quite tedius and time consuming. In such cases, normal approximation can be used with great ease and convenience. Almost all the exact sampling distribution, e.g., student's t-distribution, f-distribution, Fisher's Z-distribution and Chi-square distribution conform to normal distribution for large degrees of freedom (i.e., as n) 9.12 Glossary

Condition of homogeneity: The factors must be similar over the relevant population although, their incidence may vary from observation to observation.

Condition of independence: The factors, affecting observations, must act independently of each other.

Condition of symmetry: Various factors operate in such a way that the deviations of observations above and below mean are balanced with regard to their magnitude as well as their number.

Fitting a Normal Curve: A normal curve is fitted to the observed data with the objectives (1) To provide a visual device to judge whether it is a good fit or not. (2) Use to estimate the characteristics of the population.

Method of Areas: Under this method, the probabilities or the areas of the random variable lying in various intervals are determined. These probabilities are then multiplied by N to get the expected frequencies.

Method of Ordinates: In this method, the ordinate f(X) of the normal curve, for various values of the random variate X are obtained by using the table of ordinates for a standard normal variate.

Normal Approximation to Poisson Distribution: Normal distribution can also be used to approximate a Poisson distribution when its parameter m 10.

Normal Probability Distribution: The normal probability distribution occupies a place of central importance in Modern Statistical Theory. This distribution was first observed as the normal law of errors by the statisticians of the eighteenth century.,

9.13 Answers: Self Assessment

I. True	2. True	3. True
4. True	9. True	6. True
7. neither, nor	8.-,+	9. Normal curve
10. Method of Ordinates	11. Probabilities	12. Refer to section 9.4
13. Refer to section 9.4(c)	14. Refer to section 9.7	

9.14 Terminal Questions

%4 u Give the salient feature of a normal distribution. Write its probability function.

%4 u In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution.

%4 u A normal curve has X = 20 and10. Final the area between x, = 15 and x_2 = 40.

%4 u An aptitūde test for selecting officers in a company was conducted on 1000 candidates the average score is 42 and the standard deviation of scores is 24. assuming normal distribution for the score find:

%4 u the number of candidates whose score exceeds 60.

%4 u The number of candidates whose score lies between 30 and 60.

9.15. Suggested Readings:

%4 u Chance, W., Statistical Methods for Decision Making, Richard Irwin Inc. Home Wovel, 1969.

%4 u Feller, W., An Introduction to Probability Theory and its Applications, John Wiley & Sons Inc., New York, 1957.

%4 u Levin Richard, Statistics for Management, Prentics Hall Inc., New Delhi.

%4 u Chao Lincoin L., Statistical Methods and Analysis, McGraw Hill Inc.

%4 u Mendenhall and Reinmuth, Statistical for Management and Economics, Dexbury.

%4 u Gupta, S.P. Statistical Methods.

*****