1

DATA PRESENTATION

Unit Structure

1.0 OBJECTIVE

The learner will be able to understand variuos data types, understand frequency distributon and be able to plot simple graphs like Histograms, O give curve to display data. Also stem and leaf type of display can be learned from this chapter.

1.1 INTRODUCTION

Any Statistical study involves collecting, processing, analysing data and then reporting information from this data.

Statistics is defined as “Statistics is a science that includes the methods of collecting, organising, presenting, analysing and interpreting numerical facts and decision taken on that basis”.

1.2 DATA PRESENTATION

1.2.1 DATA TYPES

Data(or Distribution) can be classified as Ungrouped data and Grouped Data.

Grouped data can be further classified as Discrete and Continous type.

1.2.1.1 Ungrouped Data

In this type, no grouping is done on data and data is available in the raw form.

Ex 1 : Age of students in a group of five people can be 35, 38, 37, 30 and 35 years

Ex 2 : Scores of six students in a Statistics test can be 4, 6, 8, 3, 2 and 9 marks

1.2.1.2 Grouped Data

In this type data is grouped for some purpose. Grouped data can be Discrete or Continuous.

Grouped Discrete Data

Number of occurences of each discrete data can be marked as frequency of that data value in Discrete type of Data Presentation

Ex 3 : The scores of 100 students in a 10 Marks Physics class test can be grouped as :

Marks	0	1	2	3	4	5	6	7	8	9	10
Number of students	2	3	6	12	18	15	13	16	8	6	1

Ex 4: The number of students in a degree college in various courses :

Course	BCom	BMS	BScCS	BScIT	BAF
Number of students	145	98	62	48	80

Grouped Continuous Data

Some suitable class intervals are created and data is placed in the appropriate class.

Ex 5 : The scores of students in a 100 Marks Calculus class test can be grouped as :

Marks	0-40	40-60	60-75	75-100
Number of students	12	32	28	12

Ex 6: Expenses per month of families in a society are :

Expenses in Rupees	10,000-20,000	20,000-30,000	30,000-40,000	>40000
Number of families	5	12	18	3

Ex 7 : Time to manufacture an auto assembly is given in hours

Time (in hrs)	1-3	3-5	5-7	7-9	9-11
Number of assemblies	1	13	15	12	3

1.2.2 FREQUENCY DISTRIBUION

After collecting data, it can be organised in some meaningful form. The data is thus compressed in systematic manner, for example collected data can be organised in a tabular form.

Ex 8 : Following data gives marks scored by students in a test of 10 marks. Prepare frequency distribution table.
2, 4, 8, 6, 3, 4, 5, 4, 8, 6, 5, 3, 2, 0, 3, 5, 8, 9, 8, 3.

Solution:

Marks	Tally Marks	Frequency
0	|	1
1		0
2	||	2
3	||||	4
4	|||	3
5	|||	3
6	||	2
7		0
8	||||	4
9	|	1
10		0

Data can also be grouped with some suitable class Interval in frequency table.

1.2.2.1 Types of Class Intervals

Three methods of making class Intervals are :

a) Exclusive method, b) Inclusive method and c) Open end classes.

a) Exlcusive method

The upper limit of a class becomes the lower limit of the next class in this method.

For example, classes can (10-20), (20-30), (30-40) and so on.

b) Incusive method

In this type the lower limit of a class is kept onemore than the upper limit of the previsous class.

For example, classes can be (10-19), (20-29), (30-39) and so on.

a) Open end classes

In this type, the lower class limit of the first class is not given. Also the upper limit of the last class may not be given.

For example, classes can be (<100), (100-200), (200-300), (>300)

1.2.3 GRAPHS

A frequency distribution can be represented by Graphs. Graphs represent the data pictorically.

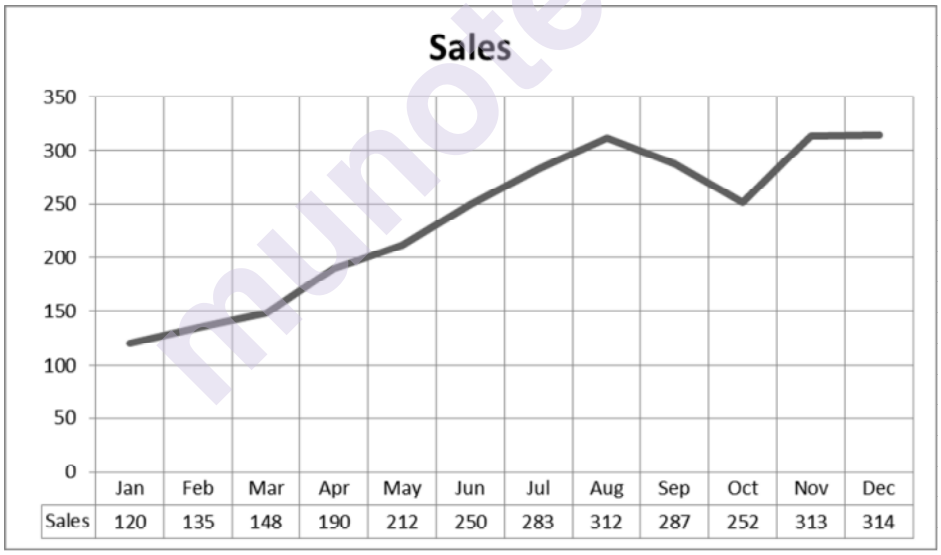
Types of Graphs :

a) Frequency curve
b) Histogram
c) O give curve
d) Stem and Leaf display

1.2.3.1 Frequency curve

Ex 9 : Plot Frequency curve

Month	Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct	Nov	Dec
Sales (in Lakh)	120	135	148	190	212	250	283	312	287	252	313	314



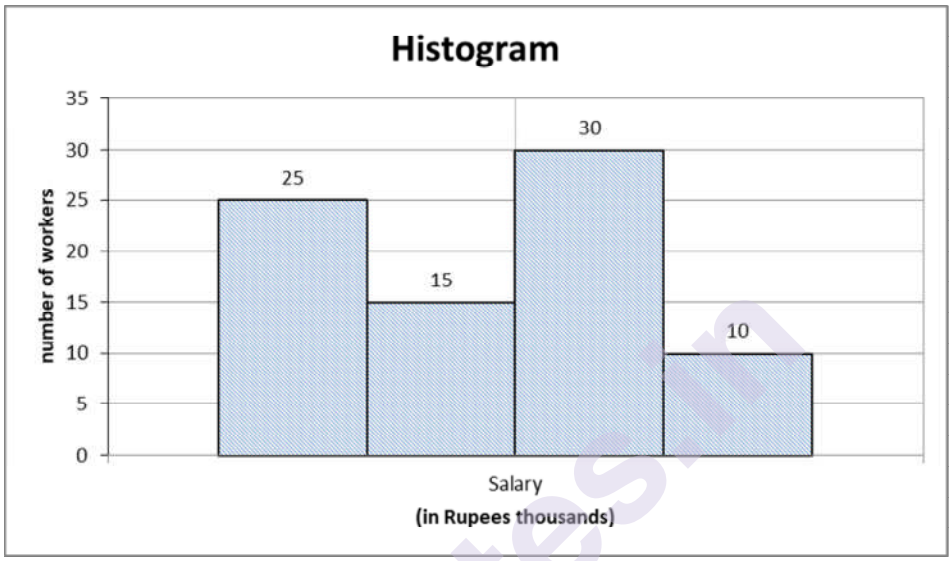
1.2.3.2 Histogram

In this type, each class is represented by a vertical bar. The bars are adjacent to each other in Histogram. The areas of the bars are proportional to the frequencies.

Ex 10 : Plot Histogram

	Number of employees
10000-20000	25
20000-30000	15
30000-40000	30
40000-50000	10

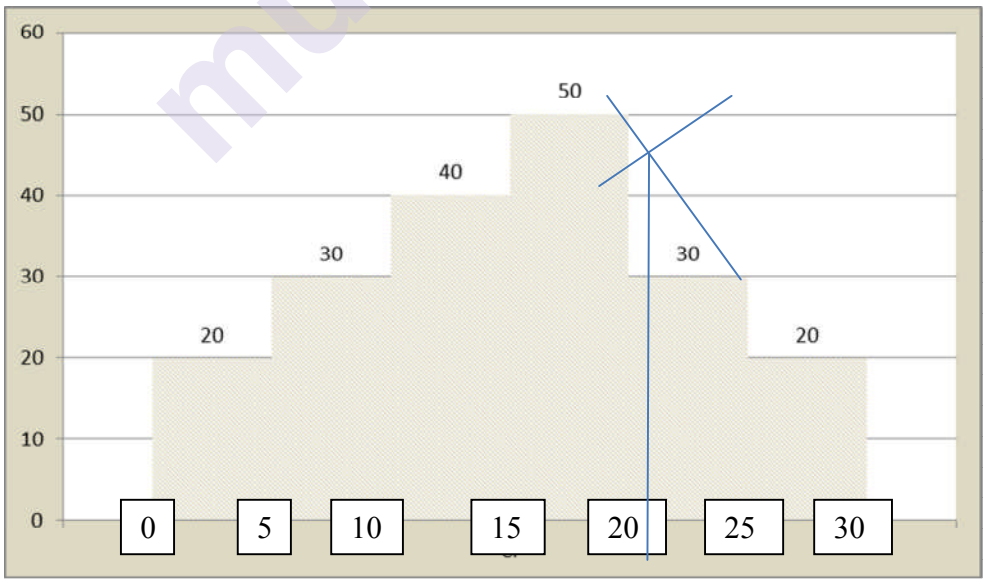
Solution :



Ex 11 : Plot Histigram and hence find Mode

CI	0-5	5-10	10-15	15-20	20-25	25-30
f	20	30	40	50	30	20

0



Mode =**15.4** **(Ans)**

1.2.3.3 O give curves

An O give curve represents the cumulative frequencies for the classes.

Ex 12 : Prepare Less than and More than cumulative frequency table.

Salary Range	No. of workers
10000-20000	125
20000-30000	134
30000-40000	150
40000-50000	85
50000-60000	15

Solution :

Salary Range	No. of workers	Less than cf	More than cf
10000-20000	125	125	510
20000-30000	134	259	385
30000-40000	150	409	251
40000-50000	85	494	101
50000-60000	16	510	16

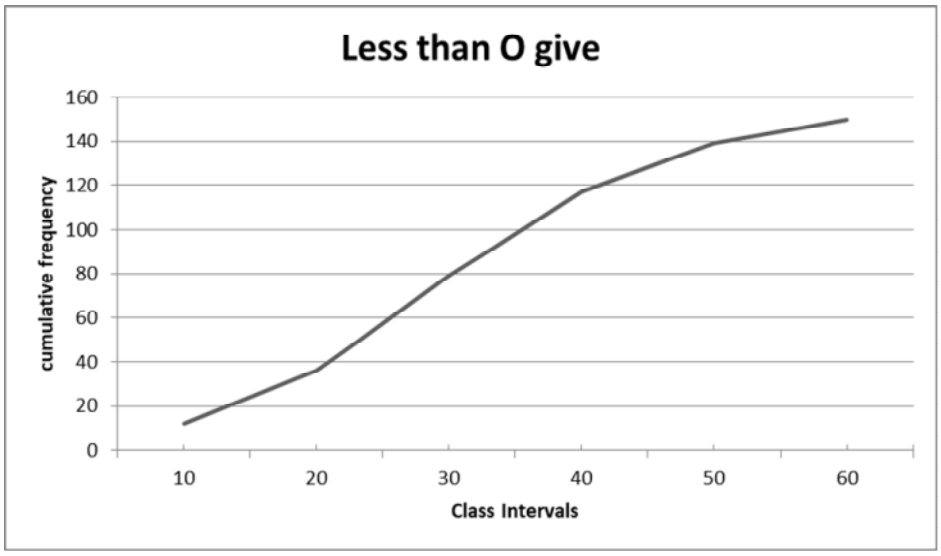
O give curves are of two types :
a) Less than O give curve and b) More than O give curve
a) Less than O give curve

Ex 13 : Plot Less than Ogive curve

Class	Frequency
10-20	12
0-30	24
30-40	43
40-50	38
50-60	22
60-70	11

Solution :

Class	Frequency	Cumulative frequency
0-10	0	0
10-20	12	12
20-30	24	36
30-40	43	79
40-50	38	117
50-60	22	139
60-70	11	150

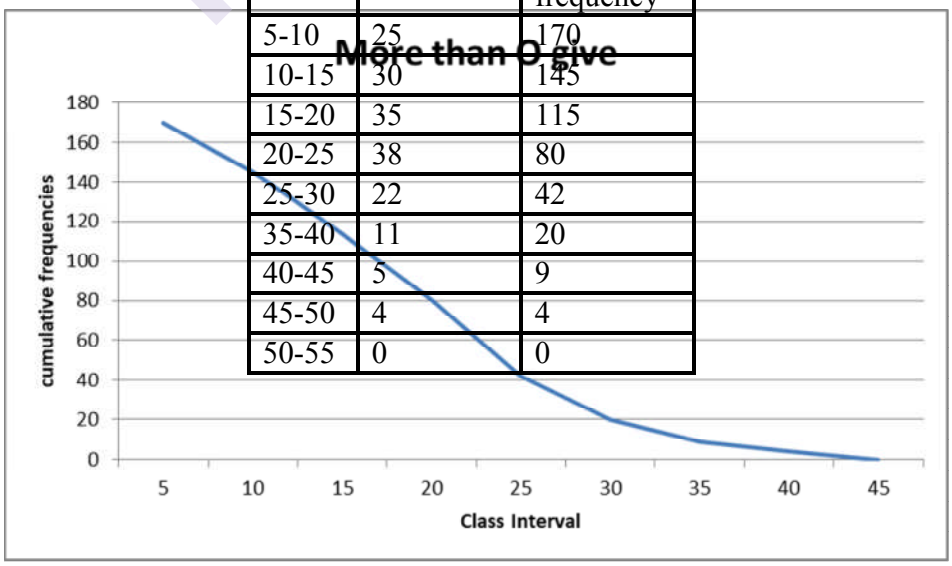


Ex 14: Plot More than Ogive curve

Class	Frequency
5-10	25
10-15	30
15-20	35
20-25	38
25-30	22
35-40	11
40-45	5
45-50	4

Solution :

Class	Frequency	More than Cumulative frequency
5-10	25	170
10-15	30	145
15-20	35	115
20-25	38	80
25-30	22	42
35-40	11	20
40-45	5	9
45-50	4	4
50-55	0	0

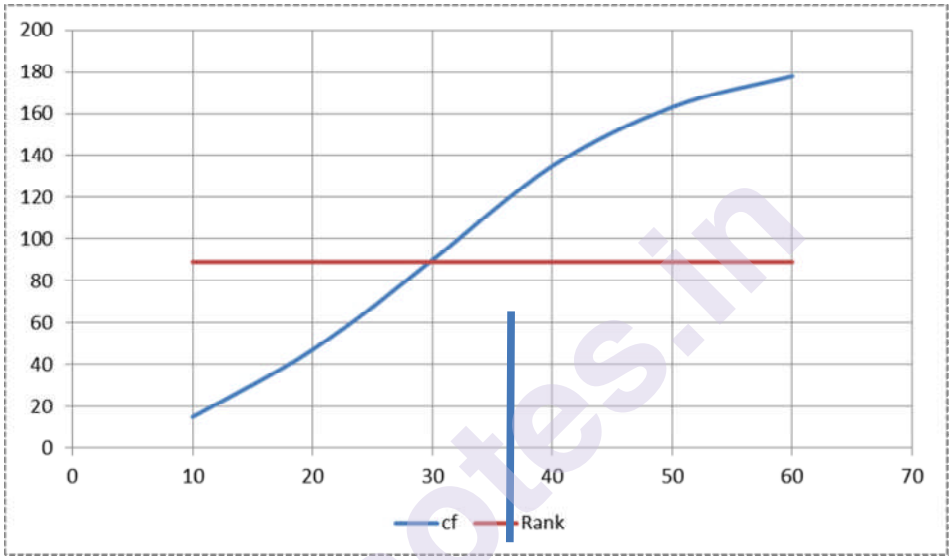


Ex 15: Plot Less than O give curve and hence find Median.

CI	0-10	10-20	20-30	30-40	40-50	50-60
f	15	32	41	45	28	15

Solution :

CI	0-10	10-20	20-30	30-40	40-50	50-60
f	15	32	43	45	28	15
Cf	15	47	90	135	163	178



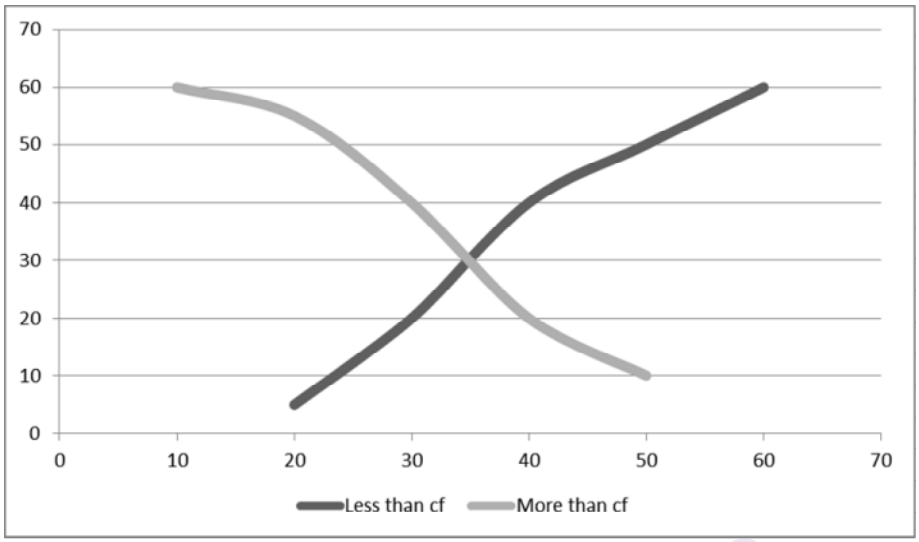
Median = **29**, the point of intersection of cf and Rank lines **Ans)**

Ex 16 : Plot Less than and More than O give curves

Range	f
10-20	5
20-30	15
30-40	20
40-50	10
50-60	10

Solution :

Range	f	Less than cf	More than cf
10-20	5	5	60
20-30	15	20	55
30-40	20	40	40
40-50	10	50	20
50-60	10	60	10



1.2.3.4 Stem and Leaf display

Stem and Leaf plot shows exact value of individual observation. It uses ungrouped data.

Steps to draw Stem and Leaf plot :

1) Divide each value of the observation into two parts. One part consisting of one or more digits as stem and rest digits as leaf.

2) The stem values are listed on the left of the vertical line and each leaf value corresponding to the stem is written in horizontal line to the right of the stem in the increasing order.

3) The stem and the leaf display gives us the ordered data and the shape of the distribution.

Ex 17 : Display the given data as stem and leaf

42, 53, 65, 63, 61, 77, 47, 56, 74, 60, 64, 68, 45, 55, 57, 82, 42, 35, 39, 51, 65, 55, 33, 76, 70, 50, 52, 54, 45, 46, 25, 36, 59, 63, 83.

Solution :

Stem	Leaf
2	5
3	3, 5, 6, 9
4	2, 2, 5, 5, 6, 7, 9
5	0, 1, 2, 3, 3, 4, 5, 5, 6, 7
6	0, 1, 3, 4, 5, 5, 8
7	0, 4, 6, 7
8	2, 3

Comparison of Histogram and Stem and Leaf plot :

1) Stem and Leaf display is simple to plot
2) Data can be easily seen in both stem and Leaf and Histogram.
3) Hsitogram is more suitable for large data set.

1.3 SUMMARY

1) Data can be of ungrouped or grouped (discrete or continuous) type
2) Frequency table gives count of observations of each variable or each class
3) Frequency curve gives data trend over period of time
4) Histogram gives pictorial representation of data in each class
5) O give curve plots cumulative frequencies in successice classs
6) Stem and Leaf plot gives more clear picture of individual data

1.4 EXERCISE

1) Explain various types of distributons with suitable examples for each.
2) Plot frequency curve

Quarter	Expenses (in K)
I	25
II	32
III	35
IV	25

3) Plot Histogram

Class	Frequency
0-4	15
4-8	22
8-12	32
12-16	25
16-20	22

4) Plot Less than O give curve

Class	Frequency
10-20	20
20-30	36
30-40	45
40-50	62
50-60	27
60-70	20

5) Plot More than O give curve

Class	Frequency
0-20	15
20-40	16
40-60	32
60-80	24
80-100	22
100-120	20

6) Draw stem and leaf plot
22, 25, 28, 32, 35, 21, 42, 42, 53, 52, 33, 35, 46, 51, 44, 34, 42, 53

7) Draw stem and leaf plot
15, 22, 26, 35, 24, 21, 25, 30, 35, 38, 24, 26, 26, 29, 32, 38, 27, 33, 35, 24, 25

1.5 LIST OFREFERENCES

1) Probability, Statistics, design of experiments and queuing theory with applications of Compter Science, S. K. Trivedi, PHI

2) Applied Statistics, S C Gupta, S Chand

❖❖❖❖

2

MEASURES OF CENTRAL TENDENCY

Unit Structure

2.0 OBJECTIVE

Learnr will be able to understand concept of Averages. Also learner will be able to take decision on correct selection of central value for the given distribution.

2.1 INTRODUCTION

It is required to convert the given set of data into some form which can represent the data. Such reduced or compressed form should be easy to interpret the distribution and also it should allow further algebraic treatment. Averages are such compact form of the distribution. Such

compact form to represent central tendency of the distribution can also be calles Averages.

Objective of a good measures of central tendency :
1) To condense the data in a single value
2) To enable comprison among various data sets

Requisites of a good Measure of Sentral tendency :
1) It should be rigidly defined.
2) It should be simple to nderstand and interpret.
3) It should cover all observations in the data set.
4) It should be capable of further algebraic treatment.
5) It should have good sampling stability.
6) It should not be undulyaffeted by extreme values.
7) It should be easy to calculate.

2.2 MEASURES OF CENTRAL TENDENCY

Types of Averages :

There are three types of Averages : Mean, Median and Mode. Also there are some more types like Geometric Mean, Harmonic Mean and Quantiles.

2.2.1 MEAN

2.2.1.1 Mean of Ungrouped Data ($\bar{x}$)

For Ungrouped Data :

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{n}$$

This can also be written as :

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n} \quad or\ simply\ \bar{x} = \frac{\sum x_i}{n}$$

Ex 1 : Find Arithmetic Mean of 4, 5, 2, 5, 7

Solution :

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{n}$$

$$\bar{x} = \frac{4 + 5 + 2 + 5 + 7}{5}$$

$$\bar{x} = 4.6 \qquad \textbf{(Ans)}$$

2.2.1.2 Mean of Grouped (Discrete) Data ($\bar{x}$)

For Grouped (discrete) Data :

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_nx_n}{f_1 + f_2 + f_3 + \cdots + f_n}$$

This can also be written as :

$$\bar{x} = \frac{\sum_{i=1}^{i=n} f_i x_i}{\sum_{i=1}^{i=n} f_i} \quad or\ simply\ \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

Ex 2 : Find Arithmetic Mean (AM) of

X	1	2	3	4	5
f	20	12	25	23	30

Solution :

X	f	fX
1	20	20
2	12	24
3	25	75
4	23	92
5	30	150
Total	**110**	**361**

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{361}{110} = 3.28$$

Mean, $\bar{x} = 3.28$ **(Ans)**

Ex 3 : Marks obtained by students of Discrete mathematics class are as given below. Find AM.

Marks	1	2	3	4	5	6	7	8	9	10
No of students	12	25	23	30	23	24	12	26	13	3

Solution :

Marks, X	1	2	3	4	5	6	7	8	9	10	**Total**
No of students, f	12	25	23	30	23	24	18	27	14	3	**191**
fX	12	50	69	120	115	144	84	208	117	30	**949**

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{949}{191} = 4.97$$

Mean, $\bar{x} = 4.97$ **(Ans)**

2.2.1.3 Mean of Grouped (Continuous) Data ($\bar{x}$)

For Grouped (continuous) Data :

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_nx_n}{f_1 + f_2 + f_3 + \cdots + f_n}$$

x_n is class mark and it is middle value of the respective class

This can also be written as :

$$\bar{x} = \frac{\sum_{i=1}^{i=n} f_i x_i}{\sum_{i=1}^{i=n} f_i} \quad \textit{or simply} \quad \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

Ex 4 : Find Arithmetic Mean (AM) of

Class Interval	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
f	4	5	11	6	5	8	9	6	4

Solution :

Class Interval	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
f	4	5	11	6	5	8	9	6	4
Class Mark, X	17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5
fX	70	112.5	302.5	195	187.5	340	427.5	315	230

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{2180}{58} = 37.59$$

Mean, $\bar{x} = 37.59$ (Ans)

Ex 5 : Find Arithmetic Mean (AM) of

Class Interval	10-20	20-30	30-40	40-50	50-60
f	15	12	18	19	21

Solution :

Class Interval	10-20	20-30	30-40	40-50	50-60	**Total**
f	15	12	18	19	21	**85**
Class Mark, x	15	25	35	45	55	
fX	225	300	630	855	1155	**3165**

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{3165}{85} = 37.24$$

Mean, $\bar{x} = 37.24$ (Ans)

2.2.1.4 Merits and Demerits of AM

Merits of AM

(i) It is rigidly defined
(ii) It is easy to calculate and easy to understand
(iii) It is based on all observations
(iv) It is capable of further algebraic treatment

Demerits of AM

(i) It is affected by extreme values
(ii) It is not possible to calculate AM for open end class intervals
(iii) It is unduly affected by extreme values
(iv) It may be number which itself may not be present in data

2.2.2 MEDIAN

2.2.2.1 Median of Ungrouped Data (M)

Median is the positional average of the data set.

Data needs to be arranged in ascending order to find the Median.

Median is middle value when there are odd number of observations.

Median is average of middle two values when there are even number of observations.

Ex 6 : Find Median of 5, 4, 3, 6, 8, 2, 5
Solution : Arrange the data in ascending order.
2, 3, 4, **5**, 5, 6, 8
Median = 5 (Ans)

Ex 7 : Find Median of 2, 4, 3, 6, 8, 2, 5, 6
Solution : Arrange the data in ascending order.
2, 2, 3, **4**, **5,** 6, 6, 8
Median = $\frac{4+5}{2} = 4.5$ (Ans)

2.2.2.2 Median of Grouped(discrete) Data (M)

Use cumulative frequency to find Median of Grouped(discrete) data.

Ex 8 : Find Median

X	1	2	3	4	5
f	20	12	25	23	30

Solution :

X	1	2	3	4	5
f	20	12	25	23	30
Cf	20	32	75	98	128

N = 128
Rank = (N+1)/2 = 129/2 = 64.5
Cf value first exceed Rank at 75. So, corresponding X value is Median
Median = 3 **(Ans)**

2.2.2.3 Median of Grouped(continuous) Data (M)

Use cumulative frequency to find Median of Grouped(continuous) data.

Steps :

1) Arrange data in ascending order

2) Obtain cumulative frequency against each class

3) Find sum of all frequencies (N).

4) Find Rank, R=N/2

5) Locate a cumulative frequency which first appears higher than Rank

6) Use given formula to find Median

$$Median, M = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$$

Where,

$l_1 = lower\ limit\ of\ Median\ class$

$l_2 = upper\ limit\ of\ Median\ class$

$R = Rank = \frac{N}{2}$

$pcf = previous\ cumulative\ frequency$

$f = frequency\ of\ Median\ class$

Ex 9 : Find Median

Class Interval	0-10	10-20	20-30	30-40	40-50
F	2	12	25	23	3

Solution :

Class Interval	0-10	10-20	**20-30**	30-40	40-50
f	2	12	**25**	23	3
Cf	2	**14**	**39**	62	65

$$R = Rank = \frac{N}{2} = \frac{65}{2} = 32.5$$

$$Median, M = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$$

$$M = 20 + \left[\frac{(32.5 - 14)(30 - 20)}{25}\right] = 20 + \left[\frac{(18.5)(10)}{25}\right] = 20 + \left[\frac{185}{25}\right] = 20 + 7.4 = 27.4$$

$\boldsymbol{M - 27.4}$ **(Ans)**

Ex 10 : Find Median

Class Interval	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
F	16	21	20	28	10	3	1	1

Solution :

Class Interval	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	Total
f	16	21	20	28	10	3	1	1	100
Cf	16	37	57	85	95	98	99	100	

$$R = Rank = \frac{N}{2} = \frac{100}{2} = 50$$

$$Median, M = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$$

$$M = 30 + \left[\frac{(50 - 37)(30 - 20)}{20}\right] = 36.50$$

$$M = 36.50$$ **(Ans)**

2.2.2.4 Merits and Demerits of MEDIAN

Merits of Median

(i) It is not affected by extreme value

(ii) It is easy to calculate. Sometimes, Median can be found out simply by observation

(iii) It can be located Graphically

(iv) It is easy to understand and easy to calulate

Demerits of Median

(i) It does not include all data in the data set

(ii) For larger data sets, arranging numbers in ascending order is tedious

(iii) It is not capable of further algebraic treatment

(iv) It does not capture small changes in data set

2.2.3MODE

Mode is the highest occuring number in the distribution, or it is the number with the highest frquency.

2.2.3.1 Mode of Ungrouped Data (Z)

Mode of ungrouped data can be simply obtained by observation. Arrange all the numbers in the ascending (or descending) order and count the occurrence of each number. The number with thehighest or most occurrence is Mode. There can bemore than Mode in the distribution.

Ex 11 : Find Mode of 7, 5, 8, 7, 6, 8, 2, 7

Solution : Arranging inascending order : 2, 8, 6, 7, 7, 7, 8, 8
Since number 7 occurred highest number of times, i.e. three times,
Mode = 7 (Ans)

Ex 12 : Find Mode of 7, 5, 8, 7, 6, 8, 2, 7, 8
Solution : Arranging inascending order : 2, 8, 6, 7, 7, 7, 8, 8, 8
Two numbres 7 and 8 bith occurred three times,
Mode = 7 and Mode = 8 (Ans)

2.2.3.2 Mode of Grouped (discrete) Data (Z)

Ex 13 : Find Mode

X	2	3	4	5	6	7	8
F	12	25	28	63	54	53	17

Since highest frequency is 63, corresponding X value is Mode.
Mode = 5 (Ans)

2.2.3.3 Mode of Grouped (continuous) Data (Z)

Following formula is to be used to find Mode of grouped (continuous) data.

$$Mode, Z = l_1 + \left[\frac{(f_1 - f_0)(l_2 - l_1)}{2f_1 - f_0 - f_2}\right]$$

Where,
$f_1 = frequency\ of\ Modal\ class$
$f_0 = frequency\ of\ class\ above\ Modal\ class$
$f_2 = frequency\ of\ class\ below\ Modal\ class$
$l_1 = lower\ limit\ of\ Modal\ class$
$l_2 = upper\ limit\ of\ Modal\ class$

Ex 14 : Find Mode

Range	0-4	4-8	8-12	**12-16**	16-20
F	12	25	28	**63**	54

Since hhighest frequency is 63, class interval [12-16] is Modal class.

$$Mode, Z = l_1 + \left[\frac{(f_1 - f_0)(l_2 - l_1)}{2f_1 - f_0 - f_2}\right]$$

$$Z = 12 + \left[\frac{(63 - 28)(16 - 12)}{2(63) - 28 - 54}\right]$$

Mode = 15.18 (Ans)

Ex 15 : Find Mode

Range	0-10	10-20	20-30	**30-40**
F	12	25	28	**63**

Since highest frequency is 63, class interval [30-40] is Modal class.

$$Mode, Z = l_1 + \left[\frac{(f_1 - f_0)(l_2 - l_1)}{2f_1 - f_0 - f_2}\right]$$

$$Z = 30 + \left[\frac{(63 - 28)(40 - 30)}{2(63) - 28 - 0}\right]$$

Mode = 33.57 (Ans)

2.2.3.4 Merits and Demerits of MODE

Merits of Mode

(i) It is not affected by extreme value

(ii) It is easy to calculate. Sometimes, Mode can be found out simply by observation

(iii) It can be located Graphically

(iv) It is easy to understand and easy to calulate

Demerits of Mode

(i) It does not include all data in the data set

(ii) Mode is not unique, hence not suitable for further algebraic treatment.

(iii) It does not capture small changes in data set

Ex 16 : The following are the weights of 30 wooden logs :

132, 166, 134, 119, 151, 114, 138, 124, 130, 132,

142, 121, 144, 147, 126, 104, 143, 129, 108, 111,

155, 131, 157, 137, 145, 122, 148, 139, 135, 136.

Arrange the data in a frequency table with class interval of 10 kg. each. The first interval being 100-110. Find Arithmetic Mean (AM), Median and Mode.

Solution :

Class Interval	Mid value (X)	Tally mark	Frequency (f)	fX	Cumulative Frequency (cf)
100-110	105	||	2	210	2
110-120	115	|||	3	345	5
120-130	125	𝍸	5	625	10
130-140	135	𝍸𝍸	10	1350	20
140-150	145	𝍸 |	6	870	26
150-160	155	|||	3	465	29
160-170	165	|	1	165	30

Mean :

$$x = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{4030}{30} = 134.33\ kg$$

Mean, $\bar{x} = 134.33$ **(Ans 1)**

Median :

$$R = Rank = \frac{N}{2} = \frac{30}{2} = 15$$

$$Median, M = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$$

$$M = 130 + \left[\frac{(15 - 10)(140 - 130)}{10}\right] = 135$$

$M = 135$ **(Ans 2)**

Mode :

$$Mode, Z = l_1 + \left[\frac{(f_1 - f_0)(l_2 - l_1)}{2f_1 - f_0 - f_2}\right]$$

$$Z = 30 + \left[\frac{(10 - 5) * 10}{2(10) - 5 - 6}\right]$$

Mode = 135.56 (Ans 3)

2.2.4 RELATIONSHIP BETWEEN MEAN, MEDIAN AND MODE

For moderately assymetrical distributions, the empirical formula relating Mean, Median and Mode is :

$$Mode = 3\ Median - 2\ Mean$$

Ex 17 : Find Mode if Mean is 12 and Median is 15

Solution :

$$Mode = 3\ Median - 2\ Mean$$

$$Mode = (3 * 15) - (2 * 12) = 45 - 24 = 21$$

$Mode = 21$ **(Ans)**

2.3 SUMMARY

Averages (Mean, Median and Mode) represent the central value in the distribution. The formula for central value depends upon the type of data. Different data sets can be compared using averages of each data set.

2.4 EXERCISE

1) Find AM of 5, 3, 2, 12, 5, 6, 9

2) Find AM of

Class Interval	0-10	10-20	20-30	30-40	40-50
f	125	123	234	220	101

3) Find Median class interval from the following distribution

X	200-202	202-204	204-206	206-208	208-210
f	145	320	445	469	342

4) Find Median

X	10	12	14	16	18
f	210	223	245	268	213

5) Find Median

X	0-4	4-8	8-12	12-16	16-20
F	65	56	43	69	34

6) Find Mode

X	6	7	8	9	10	11
F	21	23	25	37	21	15

7) Find Mode

Range	0-100	100-200	200-300	300-400	400-500
F	123	145	180	162	121

8) Find Mode if Median is 54 and Mean is 62

2.5 LIST OFREFERENCES

1) Probability, Statistics, design of experiments and queuing theory with applications of Compter Science, S. K. Trivedi, PHI

2) Applied Statistics, S C Gupta, S Chand

❖❖❖❖

3

MEASURES OF DISPERSION

Unit Structure

3.0 OBJECTIVE

The understanding of Dispersion (or deviation) is essential to completely understand and anlyse the distribution alongwith Central Tendencies. Variance, Standard Deviation and Quantiles sare useful in Data analysis. This unit helsp learner to analyse distribution using measures of deviations.

3.1 INTRODUCTION

The central value of the data can be represented by Averages, the spread of data can be exlained with the help of Measure of Dispersion.

3.2 MEASURES OF DISPERSIONS

Measure of Dispersion serve the objective of determining the reliability of an average and compare the variability of different distributions.

Requisite of a Good Measure of Dispersion :

1) It should b erogodly defined.
2) It should covr all observations in the distribution
3) It should have Sampling stability
4) It shuld be capable of further Mathematical treatment
5) It should not be duly affected by extreme values

Some important Measures of Disersion are :

1) Variance (v)
2) Standard Deviation (SD)
3) Quartile Deviation (QD)
4) Range

3.2.1 Variance

The Arithmetic Mean of squares of deviations taken from Arithmetic Mean is called **Variance**.

3.2.1.1 Variance of Ungrouped data

$$Variance, \qquad V(x) = \sigma^2 = \frac{\Sigma(x - \bar{x})^2}{n}$$

Alternate and more convinient formula for Variance is,

$$V(x) = \left[\frac{1}{n}\sum x^2\right] - (\bar{x})^2 \quad , \quad where\ \bar{x} = \frac{\Sigma x}{n}$$

Ex 1 : Find Variance of 3, 6, 8, 1, 3

$$V(x) = \left[\frac{1}{n}\sum x^2\right] - (\bar{x})^2$$

Solution :

$$V(x) = \left[\frac{119}{5}\right] - (4.2)^2$$

$$V(x) = 6.2$$ **(Ans)**

3.2.1.2 Variance of Grouped (discrete) data

$$Variance, \qquad V(x) = \sigma^2 = \frac{\Sigma f(x - \bar{x})^2}{\Sigma f}$$

Alternate and more convinient formula for Variance is,

$$V(x) = \left[\frac{\Sigma f x^2}{\Sigma f}\right] - (\bar{x})^2 \quad , \quad \text{where, } \bar{x} = \frac{\Sigma f x}{\Sigma f}$$

Ex 2 : Find Variance of

X	4	5	6	7
F	12	24	23	18

Solution :

X	4	5	6	7	Total
F	12	24	23	18	**77**
x^2	16	25	36	49	-
Fx	48	120	138	126	**432**

$$V(x) = \left[\frac{\Sigma f x^2}{\Sigma f}\right] - (x)^2$$

$$V(x) = \left[\frac{2502}{77}\right] - (5.61)^2$$

$$V(x) = 1.0$$ **(Ans)**

3.2.1.3Variance of Grouped (continuous) data

$$Variance, \quad V(x) = \sigma^2 = \frac{\Sigma f(x - \bar{x})^2}{\Sigma f}$$

Alternate and more convinient formula for Variance is,

$$V(x) - \left[\frac{\Sigma f x^2}{\Sigma f}\right] - (x)^2$$, where, $\bar{x} = \frac{\Sigma f x}{\Sigma f}$ *and x is the class mark*

Ex 3 : Find Variance of

X	0-4	4-8	8-12	12-16
F	12	24	23	18

Solution :

X	0-4	4-8	8-12	12-16	Total
f	12	24	23	18	**77**
X	2	6	10	14	-
x^2	4	36	100	196	-
Fx	48	120	138	126	**650**
fx^2	48	864	2300	3528	**6740**

$$V(x) = \left[\frac{\Sigma f x^2}{\Sigma f}\right] - (x)^2$$

$$V(x) = \left[\frac{6740}{77}\right] - (8.44)^2$$

$$V(x) = 16.3$$ **(Ans)**

3.2.2 Standard Deviation

Standard Deviation is square root of the variance. One can find variance and then take square root of variance, which will give standard deviation

3.2.2.1 Standard Deviation of Ungrouped data

$$sd = \sqrt{\left[\frac{\Sigma x^2}{n}\right] - (x)^2}$$

Ex 4 : Find standard deviation of 3, 6, 8, 1, 3

Solution :

$$x = \frac{\Sigma fx}{\Sigma f} = 4.2$$

$$sd = \sqrt{\left[\frac{\Sigma x^2}{n}\right] - (x)^2}$$

$$sd = \sqrt{\left[\frac{119}{5}\right] - (4.2)^2}$$

$$sd = \sqrt{6.2}$$

$$sd = 2.48$$ **(Ans)**

Ex 5 : Find standard deviation of **49, 63, 46, 59, 65, 52, 60, 54**

$$\bar{x} = \frac{\Sigma x}{n} = \frac{448}{8} = 56.00$$

$$sd = \sqrt{\left[\frac{\Sigma x^2}{n}\right] - (x)^2}$$

$$sd = \sqrt{\left[\frac{25412}{8}\right] - (56)^2}$$

$$sd = \sqrt{40.5}$$

$$sd = 6.36$$ **(Ans)**

3.2.2.2 Standard Deviation of Grouped (discrete) data

Standard deviation of Grouped (discrete) data can be found out by taking square root of variance

$$sd = \sqrt{\left[\frac{\Sigma fx^2}{\Sigma f}\right] - (x)^2}$$

Ex 6 : Find Standard Deviation

X	2	3	4	5	6	7	8	9
f	2	3	4	2	5	3	2	1

Solution :

x	2	3	4	5	6	7	8	9	**Total**
f	2	3	4	2	5	3	2	1	**22**
fx	4	9	16	10	30	21	16	9	**115**
fx^2	8	27	64	50	180	147	128	81	**685**

$$sd - \sqrt{\left[\frac{\Sigma fx^2}{\Sigma f}\right] - (x)^2}$$

$$sd = \sqrt{\left[\frac{685}{22}\right] - (5.23)^2}$$

$$sd = \sqrt{3.8} = 1.95$$

$$sd = 1.95$$ **(Ans)**

3.2.2.3 Standard Deviation of Grouped (continuous) data

Standard deviation of Grouped (continuous) data can be found out by taking square root of variance

$$sd = \sqrt{\left[\frac{\Sigma f x^2}{\Sigma f}\right] - (x)^2}$$

Ex 7 : Find standard deviation

X	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	Total
F	2	5	3	6	4	2	1	1	
fX									

3.2.2.4Combined Mean and combined Standard Deviation

Combined Mean :

Combined Mean of two data sets can be found out using following formula.

$$\textit{Combined Mean,} \quad x = \frac{n_1 x_1 + n_2 x_2}{n_1 + n_2}$$

Ex 8 : Find combined Mean of following data sets.

	Set 1	Set 2
Number of observations	25	45
Mean	8	9

Solution :

$$\textit{Combined Mean,} \quad x = \frac{n_1 x_1 + n_2 x_2}{n_1 + n_2}$$

$$\textit{Combined Mean,} \quad x = \frac{(25)(8) + (45)(9)}{25 + 45} = \frac{200 + 405}{70} = \frac{605}{70} = 8.64$$

$$\textit{Combined Mean} = 8.64 \qquad \textbf{(Ans)}$$

Ex 9 : Find Combined Mean

	Set 1	Set 2	Set 3
Number of observations	120	135	145
Mean	51	48	46

Solution :

$$\textit{Combined Mean,} \quad x = \frac{n_1 x_1 + n_2 x_2 + n_3 x_3}{n_1 + n_2 + n_3}$$

$$\textit{Combined Mean,} \quad x = \frac{(120)(51) + (135)(48) + (145)(46)}{51 + 40 + 46} = 48.17$$

$$\textit{Combined Mean} = 48.17 \qquad \textbf{(Ans)}$$

Combined Standard Deviation :

Combined Standard Deviation, $\sigma = \sqrt{\frac{n_1[(\sigma]_1^2 + d_1^2) + n_2[(\sigma]_2^2 + d_2^2)}{[(n]_1 + n_2)}}$

Where,

$d_1^2 = (\bar{x} - \overline{x_1})^2$ *and* $d_2^2 = (\bar{x} - \overline{x_2})^2$

Ex 10 : Find Combined Mean and Combined Standard Deviation :

	Group 1	Group 2
No. of observations	32	25
Mean	12	14
SD	3	4

Solution :

	Group 1	Group 2
No. of observations	$n_1 = 32$	$n_2 = 25$
Mean	$\overline{x}_1 = 12$	$\overline{x}_2 = 14$
SD	$sd_1 = 3$ $sd_1^2 = 9$	$sd_2 = 4$ $sd_2^2 = 16$

Combined Mean, $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

Combined Mean, $\bar{x} = \frac{(32)(12) + (25)(14)}{32 + 25} = \frac{384 + 350}{57} = \frac{734}{57} = 12.87$

Combined Mean $= 12.87$

$d_1^2 = (\bar{x} - \overline{x_1})^2$ *and* $d_2^2 = (\bar{x} - \overline{x_2})^2$

$d_1^2 = (12.87 - 12)^2$ *and* $d_2^2 = (12.87 - 14)^2$

$d_1^2 = 0.76$ *and* $d_2^2 = 1.26$

Combined Standard Deviation, $\sigma = \sqrt{\frac{n_1[(\sigma]_1^2 + d_1^2) + n_2[(\sigma]_2^2 + d_2^2)}{[(n]_1 + n_2)}}$

Combined Standard Deviation, $\sigma = \sqrt{\frac{(32)(9 + 0.76) + (25)(16 + 1.26)}{(32 + 25)}}$

Combined Standard Deviation, $\sigma = 3.61$ **(Ans)**

3.2.3 Coefficient of Variation (CV)

The Coefficient of Variation is the ratio of standard deviation to the arithmetic mean expressed as percentage.

$CV = \frac{\sigma}{\bar{x}}$ 100

CV can be used to know the consistency of the data. A distribution with smaller CV is more consistent than the other one. CV is also useful for comparing two or more sets of data that are measued in different units of measurement.

Ex 11 : Find coefficient of variation of 2, 5, 4, 1 and 3

Solution : $sd = \sqrt{\left[\frac{1}{n}\sum x^2\right] - (x)^2}$

$sd = \sqrt{\left[\frac{55}{5}\right] - (3)^2}$

$sd = \sqrt{2}$

$sd = 1.41$

$CV = \frac{\sigma}{x}\ 100$

$CV = \frac{1.41}{3}\ 100 = 0.47$ x 100 = **47%** **(Ans)**

3.2.4 Quartile Deviation (QD)

Quartile Deviation is defined as ,

$QD = \frac{Q_3 - Q_1}{2}$

Where, Q3 is upper (third) quartil and Q1 is lower (first) quartile.

Q_i is defined as,

$Q_i = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$, where $R = \frac{i}{4} * N$ for $i = 1, 2\ or\ 3$

Coefficient of QD is defined as,

$QD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Ex 12 : Find QD

Class Interval	0-10	10-20	20-30	30-40	40-50
f	2	12	25	23	3

Solution :

Class Interval	0-10	10-20	20-30	30-40	40-50
f	2	12	25	23	3
Cf	2	14	39	62	65

To find Q3 :

$R = Rank = \frac{3}{4}N = \frac{3}{4} * 65 = 48.75$

Select cumulative frequency value higher or equal to Rank,

$Q_3 = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$

$Q_3 = 30 + \left[\frac{(48.75 - 39)(40 - 30)}{23}\right] = 30 + \left[\frac{(9.75)(10)}{23}\right] = 30 + \left[\frac{97.5}{23}\right] = 30 + 4.23$

$Q_3 = 34.23$

To find Q1 :

$$R = Rank = \frac{1}{4}N = \frac{1}{4} * 65 = 16.25$$

Select cumulative frequency value higher or equal to Rank,

$$Q_1 = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$$

$$Q_1 = 20 + \left[\frac{(16.25 - 14)(30 - 20)}{25}\right] = 20 + \left[\frac{(2.25)(10)}{25}\right] = 20 + \left[\frac{22.5}{25}\right] = 20 + 0.9$$

$$\boldsymbol{Q_1 = 20.9}$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{34.23 - 20.9}{2} = 6.69$$

$$\boldsymbol{QD = 6.69}$$ **(Ans)**

Ex 13 : Find Co-efficient of QD

Class Interval	0-2	2-4	4-6	6-8	8-10
f	14	18	21	20	12

Solution :

Class Interval	0-2	2-4	4-6	6-8	8-10
f	14	18	21	20	12
Cf	14	32	53	73	85

To find Q3 :

$$R = Rank = \frac{3}{4}N = \frac{3}{4} * 85 = 63.75$$

Select cumulative frequency value higher or equal to Rank,

$$Q_3 = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$$

$$Q_3 = 6 + \left[\frac{(63.75 - 53)(8 - 6)}{20}\right] = 7.08$$

$$\boldsymbol{Q_3 = 7.08}$$

To find Q1 :

$$R = Rank = \frac{1}{4}N = \frac{1}{4} * 85 = 21.25$$

Select cumulative frequency value higher or equal to Rank,

$$Q_1 = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$$

$$Q_1 = 2 + \left[\frac{(21.25 - 14)(4 - 2)}{18}\right] = 2.81$$

$$Q_1 = 2.81$$

$$QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{7.08 - 2.81}{7.08 + 2.81} = 0.43$$

$$Co - efficient\ of\ QD = 0.43$$ **(Ans)**

Merits and Demerits of QD

Merits of QD :

1) It is rigidly defined
2) It is not affected by extreme values
3) It can be calculated with open end class intervals

Demerits of QD :

1) It is not based on alll observations
2) It is much affected by sampling fluctuations

3.3 SUMMARY

1) Standard Deviation and Variance are two important measures of Dispersion.
2) Coefficient of Variation is the ration of standard deviation to mean expressed as percentage.

3.4 EXERCISE

1) Find SD of 4, 6, 2, 8, 2

2) Find Variance of

X	2	3	4	5	6
F	65	78	110	88	86

3) Find Standard Deviation of

Range	10-20	20-30	30-40	40-50	50-60	60-70	70-80
F	5	4	8	9	4	5	3

4) Find QD and Coefficient of QD of

Range	0-4	4-8	8-12	12-16	16-20	20-24	24-28
F	5	12	24	18	16	12	1

5) Find Combined Mean and Combined Standard Deviation

	Group 1	Group 2	Group 3
No. of observations	120	135	130
Mean	13	16	15
SD	3	5	4

3.5 LIST OFREFERENCES

1) Probability, Statistics, design of experiments and queuing theory with applications of Compter Science, S. K. Trivedi, PHI

2) Applied Statistics, S C Gupta, S Chand

❖❖❖❖

4

MOMENTS, SKEWNESS AND KURTOSIS

Unit Structure

4.0 OBJECTIVE

Moments are used to describe characteristics of a distribution such as central tendency, dispersion. Skewness refers to the lack of symmetry of the curve on both sides, whereas, Kurtosis referes to peakedness of the normal distribution curve.

4.1 INTRODUCTION

Moments are a family of equations, each representing a different quantity.

Skewness refers to lack of symmetry in the distribution, whereas Kurtosis refers to peakedness of the normal distribution curve.

Skewness is represented by either Karl Pearson's measure or Bowley's measure of Skewness.

4.2 MOMENTS

Moments can be defined as arithmetic mean of different powers of deviations of observations from a particular value. When that particlular value is zero, moment is called **raw moment**, and when that value is mean, moment is called **central moment**.

For ungrouped data :

If $x_1, x_2, \ldots, x_n$ are data values, then Raw Moment is given as :

$$\mu'_r = \frac{\sum x_i^r}{n}, \qquad where\ r = 0, 1, 2, \ldots$$

Central Moment for ungrouped data is given as :

$$\mu_r = \frac{\sum (x_i - \bar{x})^r}{n}, \qquad where\ r = 0, 1, 2, \ldots$$

In general Moment around a point a is given as

$$\mu_r = \frac{\sum (x_i - a)^r}{n}, \qquad where\ r = 0, 1, 2, \ldots$$

For Grouped data :

If $x_1, x_2, \ldots, x_n$ are data values(or class marks) with corresponding frequency values as $f_1, f_2, \ldots, f_n$, then , then Raw Moment is given as :

$$\mu'_r = \frac{\sum f_i x_i^r}{\sum f_i}, \qquad where\ r = 0, 1, 2, \ldots$$

Central Moment for grouped data is given as :

$$\mu_{(r,\bar{x})} = \frac{\sum f_i (x_i - \bar{x})^r}{\sum f_i}, \qquad where\ r = 0, 1, 2, \ldots$$

In general Moment around a point a is given as

$$\mu_{(r,a)} = \frac{\sum f_i (x_i - a)^r}{\sum f_i}, \qquad where\ r = 0, 1, 2, \ldots$$

Ex 1: Find first four raw moments of following data :

X	2	3	4	5
f	12	15	18	15

Solution :

	X	f	fX	fX^2	fX^3	fX^4
	2	12	24	48	96	192
	3	15	45	135	405	1215
	4	18	72	288	1152	4608
	5	15	75	375	1875	9375
Total	**14**	**60**	**216**	**846**	**3528**	**15390**

Raw Moments, $\mu'_r = \frac{\sum f_i x_i^r}{\sum f_i}$

First Raw Moment : $\mu'_1 = \frac{216}{60} = 3.6$

Second Raw Moment : $\mu_2' = \frac{846}{60} - 14.1$

Third Raw Moment : $\mu_3' = \frac{3528}{60} = 58.8$

Fourth Raw Moment : $\mu_4' = \frac{15390}{60} = 256.5$ **(Ans)**

4.3 RELATION BETWEEN CENTRAL MOMENTS AND RAW MOMENTS

$\mu_r' = \frac{\sum x_i^r}{n}$, $where\ r = 0, 1, 2, \ldots$

$for\ r = 0, \mu_0' = \frac{\Sigma x_i^0}{n} = 1$

$for\ r = 0,\ \mu_0 = \frac{\Sigma (x_i - \bar{x})^0}{n} = 1$

$for\ r = 1, \quad \mu_1' = \frac{\sum x_i^1}{n} = \bar{x}$

$for\ r = 1,\ \mu_1 = \frac{\sum (x_i - \bar{x})^1}{n} = 0,$ *since the sum of deviations from mean is zero*

$for\ r = 2, \quad \mu_2 = \mu_2' - ([\mu_1'])^2$

$for\ r = 3, \quad \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2([\mu_1'])^3$

$for\ r = 4, \quad \mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'([\mu_1'])^2 - 3([\mu_1'])^4$

For grouped data, these results can be proved by replacing $\sum x$ *with* $\sum fx$

4.4 SKEWNESS

Skewness refers to deviation from (or lack of) symmetry. A curve which is not symmetric about any central vlaue on both the sides is called skewed curve. When data is perfectly symmetrical about both the sides, mean, median and mode coinicide at the central point. In case of skewness, they change their position relative to each other.

Skewness can positive or negative.

Skewness measurement can be Absolute or Relative.

Absolute measures of Skewness :

There are two absolute measure**s.**

1) Karl Pearson's measure of Skewness = Mean - Mode

2) Bowley's measure of Skewness = $(Q_3 - Q_2) - (Q_2 - Q_1)$,

Where,

Q_1, Q_2 and Q_3 are first, second and third quartiles respectively

Relative measures of Skewness :

There are three relative measures of Skewness.

1) **Karl Pearson's coefficient of Skewness,** $SK_p = \frac{Mean - Mode}{sd}$

$If\ SK_p > 0, it\ is\ positively\ skewed\ curve$

$If\ SK_p = 0, it\ is\ symmetric\ curve$

$If\ SK_p < 0, it\ is\ negatively\ skewed\ curve$

2) **Bowley's coefficient of Skewness,** $SK_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{[(Q]_3 + Q_1 - 2Q_2)}{[(Q]_3 - Q_1)}$

$If\ SK_B > 0, it\ is\ positively\ skewed\ curve$

$If\ SK_B = 0, it\ is\ symmetric\ curve$

$If\ SK_B < 0, it\ is\ negatively\ skewed\ curve$

Bowley's coefficient of Skewness lies between -1 to +1

3) **Relative measure based on Moments,** $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

$\gamma_1 = \pm\beta_1$, sign of γ_1 depends upon the sign of μ_3

$If\ \gamma_1 > 0, it\ is\ positively\ skewed\ curve$

$If\ \gamma_1 = 0, it\ is\ symmetric\ curve$

$If\ \gamma_1 < 0, it\ is\ negatively\ skewed\ curve$

Ex 2: Find Karl Pearson's coefficient of Skewness for 4, 5, 3, 5, 5

Solution : Mean = $\frac{\Sigma x}{n} = \frac{22}{5} = 4.4$

Mode = **5**

$$sd = \sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2} = \sqrt{\frac{100}{5} - (4.4)^2} = 0.8$$

$$SK_p = \frac{Mean - Mode}{sd} = \frac{4.4 - 5}{0.8} = -0.75$$ (negavive skewness)

(Ans)

Ex 3: Find Bowley's coefficient of Skewness for the following data.

Score	0-20	20-40	40-60	60-80	80-100
Number of student	15	25	32	35	16

Solution :

Score	0-20	20-40	40-60	60-80	80-100
Number of student	15	25	32	35	16
cf	15	40	72	107	123

Bowley's measure of Skewness, $SK_B = \frac{[(Q]_3 + Q_1 - 2Q_2)}{[(Q]_3 - Q_1)}$

<u>To find Q1 :</u>

$R = Rank = \frac{1}{4}N = \frac{1}{4} * 123 = 30.75$

Select cumulative frequency value higher or equal to Rank,

$Q_1 = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$

$Q_1 = 20 + \left[\frac{(30.75 - 15)(40 - 20)}{25}\right] = 32.60$

$\boldsymbol{Q_1 = 32.60}$

<u>To find Q2 :</u>

$R = Rank = \frac{2}{4}N = \frac{1}{2} * 123 = 61.50$

Select cumulative frequency value higher or equal to Rank,

$Q_2 = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$

$Q_2 = 40 + \left[\frac{(61.50 - 40)(60 - 40)}{32}\right] = 53.44$

$\boldsymbol{Q_2 = 53.44}$

<u>To find Q3 :</u>

$R = Rank = \frac{3}{4}N = \frac{3}{4} * 123 = 92.25$

Select cumulative frequency value higher or equal to Rank,

$Q_3 = l_1 + \left[\frac{(R - pcf)(l_2 - l_1)}{f}\right]$

$Q_3 = 60 + \left[\frac{(92.25 - 72)(80 - 60)}{35}\right] = 71.57$

$\boldsymbol{Q_3 = 71.57}$

Bowley's measure of Skewness, $SK_B = \frac{[(Q]_3 + Q_1 - 2Q_2)}{[(Q]_3 - Q_1)}$

$SK_B = \frac{(71.57 + 32.60 - 2 * 53.44)}{(71.57 - 32.60)}$

$SK_B = -0.07$, slight negative Skewness **(Ans)**

Ex 4 : Find Karl Pearson's coefficient of Skewness

Range	f
20-40	15
40-60	20
60-80	35
80-100	12
100-120	5

Solution :

Range	F	X	fX		X^2	fX^2
20-40	15	30	450		900	13500
40-60	20	50	1000	f_0	2500	50000
60-80	35	70	2450	f_1	4900	171500
80-100	12	90	1080	f_2	8100	97200
100-120	5	110	550		12100	60500
Total	**87**		**5530**		**28500**	**392700**

$$Mean, \quad \bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{5530}{87} = 63.56$$

$$Mode, \quad Z = l1 + \frac{(f_1 - f_0)(l_2 - l_1)}{2f_1 - f_0 - f_2} = 60 + \frac{(35-20)(80-60)}{2*35-20-12} = 67.89$$

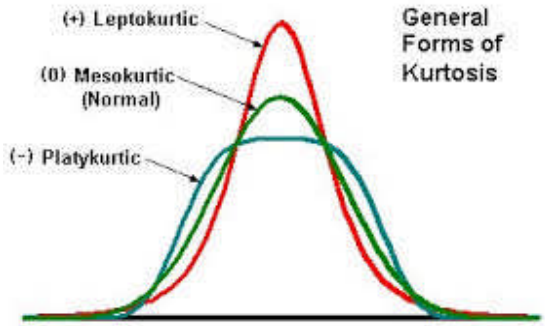
$$sd = \sqrt{\frac{\Sigma fX^2}{\Sigma f} - \left(\frac{\Sigma fX}{\Sigma f}\right)^2} = \sqrt{\frac{392700}{87} - \left(\frac{5530}{87}\right)^2} = 21.67$$

$$\textbf{Karl Pearson's coefficient of Skewness}, \quad SK_p = \frac{63.56 - 67.89}{21.67} = -0.20$$

The curve is slightly negatively skewed **(Ans)**

4.5 KURTOSIS

Normal distrbution curve is bell shaped in nature. But two distribution may have symmetry, but their peakedness may vary. One may have more height than the other. This characteristic is known as**Kurtosis**. The main reason for this variation in peak is concentration of data around the mean value. The curve will have higher peak for smaller standard deviation.



A distribution that is peaked in the same way as any normal distribution is termed as **Mesokurtic**.

A **Leptokurtic** distribution is one with higher peak compared to Mesokurtic distribution. The curne has higher peak and is thin.

In contrast to Leptokurtic distribution, **Platykurtic** distribution is flattened from top and has broad appearance compared to Mesokurtic curves.

Measure of Kurtosis :

$Measure\ of\ Kurtosis, \quad \beta_2 = \frac{\mu_4}{\mu_2^2}$

$and, \quad \gamma_2 = \beta_2 - 3$

For Mesokurtic distribution, $\beta_2 = 3$, and $\gamma_2 = 0$

For Leptokurtic distribution, $\beta_2 > 3$, and $\gamma_2 > 0$

For Platykurtic distribution, $\beta_2 < 3$, and $\gamma_2 < 0$

Both β_2 *and* γ_2 are unit free parameters and are independent of change of scale and change of origin.

4.6 SUMMARY

1) Moments describe various parameters

2) Raw moments and Central moments can be related with various formulas

3) Skewness represent extent of lack of symmetry in un symmetrical distributions

4) Karl Pearson's measure of Skewness and Bowley's co efficient of Skewness are measures of Skewness

5) Kurtosis represent thinness or flattened but symmetrical normal distribution curves

6) Kurtosis can be Mesokurtic, Laptokurtic or Platykurtic

4.7 EXERCISE

1) Expian Karl Pearson's co-efficient of Skewness.

2) Find Karl Pearson's coefficient of Skewness for 12, 14, 13, 16, 18

3) Find Bowley's coefficient of Skewness for the following data.

Score	0-10	10-20	20-30	30-40	40-50
Number of student	23	42	45	40	12

4) Given $\mu_4 = 1024,$ and $\mu_2 = 16,$ find γ_2

4.8 LIST OFREFERENCES

1) Probability, Statistics, design of experiments and queuing theory with applications of Compter Science, S. K. Trivedi, PHI

2) Applied Statistics, S C Gupta, S Chand

5

CORRELATION AND REGRESSION ANALYSIS

Unit Structure

5.0 OBJECTIVE

Correlation, as name suggests correlates two parameters. Statistically, Correlation coefficient gives an estimate of extent of correlation between these two parameters (or quantities). One can correlate score in final exam with the number of hours of study during the term.

Regression is an estimation technique. It uses historical data to estimate the possible value of that parameter in future. Regression analysis helps to allocate resources based on estimation of the parameter like estimation of future sales or estimation of future climatic condition.

5.1 INTRODUCTION

Correlation can be measured statistically by Coefficient of Correlation or even Scatter graph can be used.

Regression equation can be obtained either by method of least squares or one can even use Regression coefficient.

5.2 CORRELATION

Correlation analysis provides information about changes in one parameter with reference to changes in othe rparameter. When one variable increases, the other also increases (may be in different extent), then the correlation is positive. In contrast to this, when variable increases, the other dcreases, the correlation can be termed negative. There can instances when there is no correlation between two parameters.

Correlation can be represented by :

1) Scatter Graph (Graphical representation) or
2) Karl Pearson's coefficient of correlation (r) which is a stastical measure of correlation

5.2.1 SCATTER GRAPH

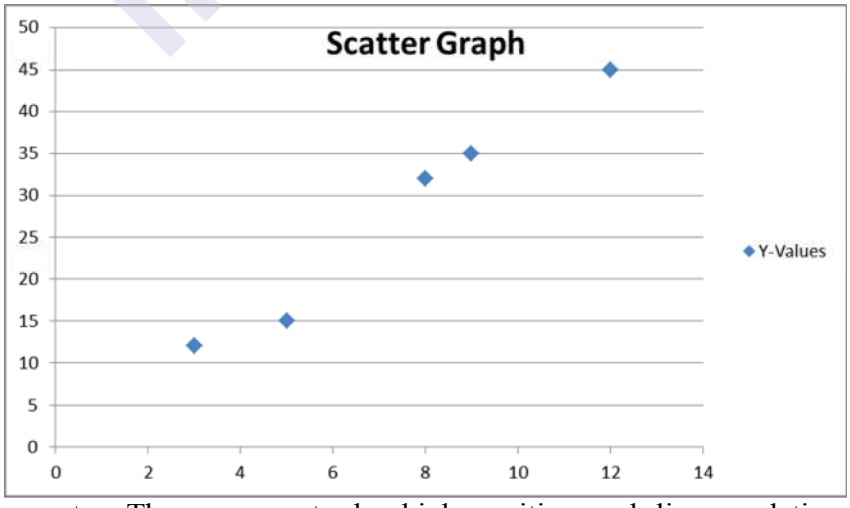
Scatter Graph, also called X-Y plot gives following information about two paratemers :

1) Shape (linear or non linear)
2) Extent of correlation
3) Nature of correlation like positive, negative or no correlation

Ex 1 : Plot Scatter Graph and comment.

X	Y
3	12
5	15
8	32
9	35
12	45

Solution :

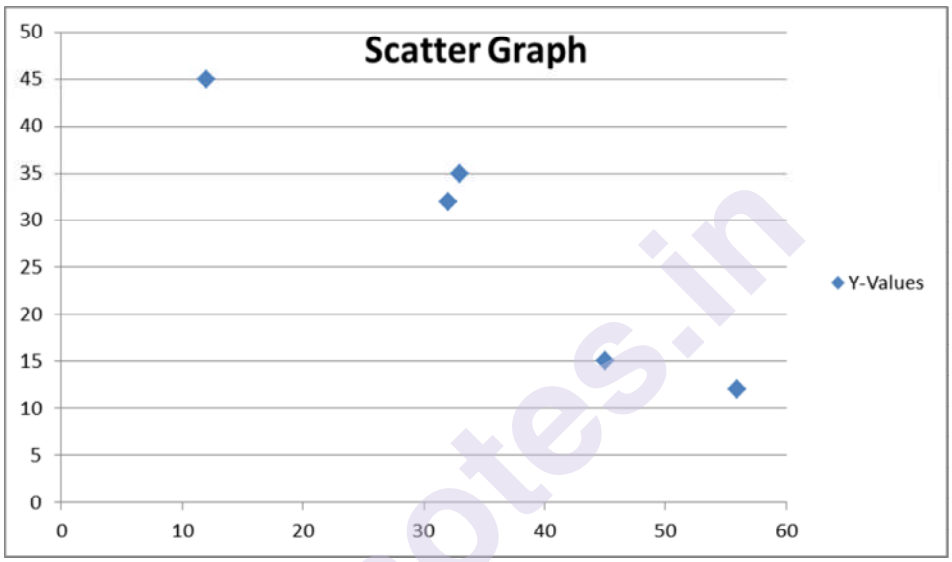


Comment : There seems to be high positive and linear relationship between X and Y

(Ans)

Ex 2 : Plot Scatter Graph and comment.

X	Y
56	12
45	15
32	32
22	35
12	45



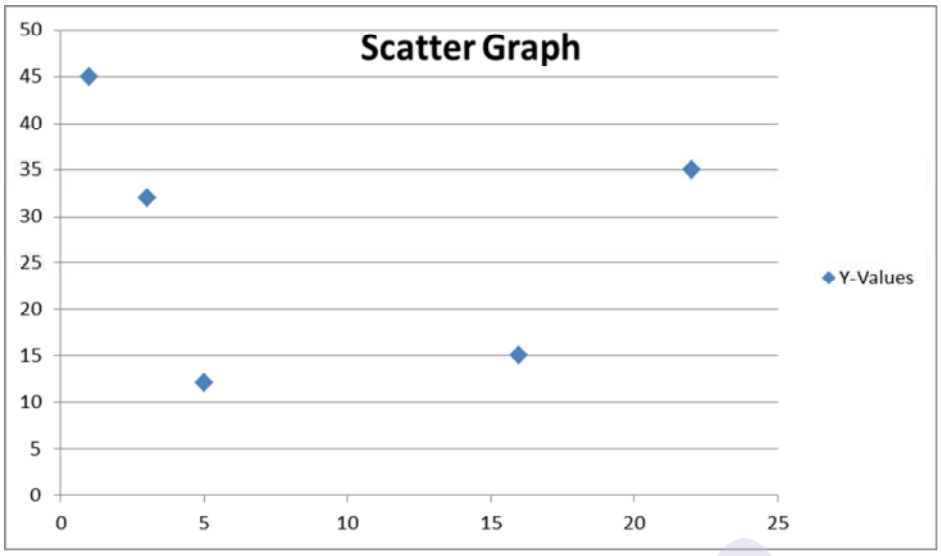
Comment : There seems to be high negative and linear relationship between X and Y

(Ans)

Ex 3 : Plot Scatter Graph and comment.

X	Y
5	12
16	15
3	32
22	35
1	45

Solution :



Comment : There seems to be slight negative or no correlation between X and Y

(Ans)

Merits and Demerits of Scatter Graph

Merits :

1) Scatter Graph is easy to plot
2) It is also easy to understand and interpret general trend
3) Non linear relation can be easily detected
4) Scatter graph can very easily spot some abnormal values which are ot consistent with rest of the values

Demerits :

1) Scatter graph does not give mathematical (or numerical) value of the correlation, hence can not be used in further calculations, except for visual observations
2) This method is useful for relatively small number of observations
3) It can not be applied to qualitative data whose numberical values are not available like emotions, sentimets correlation can not be represented by Scatter Graph as no numerical values are available

5.2.2 KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl Pearson's coefficient of correlation (r) is used to find tpe of correlation i.e. positive, negative or no correlation and also extent of correlation like strong, medium or weak correlation.

It is a numerical measure of correlation and is very useful in statistical analysis.

Basic definition of **r** is $$r = \frac{cov(X,Y)}{sd_x sd_y}$$

But, working formula for **r** is,

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\ \Sigma x^2 - (\Sigma x)^2}\sqrt{n\ \Sigma y^2 - (\Sigma y)^2}}$$

Ex 4 : Find Karl Pearson's coefficient of correlation

X	Y
3	12
5	15
8	32
9	35
12	45

Solution :

	X	**Y**	**XY**	**X²**	**Y²**
	3	12	36	9	144
	5	15	75	25	225
	8	32	256	64	1024
	9	35	315	81	1225
	12	45	540	144	2025
Total	**37**	**139**	**1222**	**323**	**4643**

n = 5, number of ordered pairs

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\ \Sigma x^2 - (\Sigma x)^2}\sqrt{n\ \Sigma y^2 - (\Sigma y)^2}}$$

$r = ((5)(1222) - (37)(139))/(\sqrt{((5)(323 - [(37)]^2)}\sqrt{((5)(4643 - [(139)]^2))} = 0.9880$

$r = 0.9880$

There is very strong positive correlation between X and Y **(Ans)**

Ex 5 : Find Karl Pearson's coefficient of correlation

X	Y
56	12
45	15
32	32
22	35
12	45

Solution :

	X	Y	XY	X^2	Y^2
	56	12	672	3136	144
	45	15	675	2025	225
	32	32	1024	1024	1024
	22	35	770	484	1225
	12	45	540	144	2025
Total	**37**	**139**	**1222**	**323**	**4643**

n = 5, number of ordered pairs

$$r = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\,\Sigma y^2 - (\Sigma y)^2}}$$

$r = ((5)(3681) - (167)(139))/(\sqrt{((5)(6813) - [(167)]^{\wedge}2)}\sqrt{((5)(4643 - [(139)]^{\wedge}2))} = -0.9804$

$$r = -0.9804$$

There is very strong negative correlation between X and Y **(Ans)**

Ex 6 : Find Karl Pearson's coefficient of correlation

X	Y
5	12
16	15
3	32
22	35
1	45

Solution :

	X	Y	XY	X^2	Y^2
	5	12	60	25	144
	16	15	240	256	225
	3	32	96	9	1024
	22	35	770	484	1225
	1	45	45	1	2025
Total	**47**	**139**	**1211**	**775**	**4643**

n = 5, number of ordered pairs

$$r = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\,\Sigma y^2 - (\Sigma y)^2}}$$

$r = ((5)(1211) - (47)(139))/(\sqrt{((5)(775) - [(47)]^{\wedge}2)}\sqrt{((5)(4643 - [(139)]^{\wedge}2))} = -0.1877$

$$r = -0.1877$$

There is slight negative correlation between X and Y **(Ans)**

5.2.3 PROPERTIES OF KARL PEARSON'S COEFFICIENT OF CORRELATION

1) Correlation coefficient lies between -1 and +1
2) Correlation coefficient is independent of change of origin and scale
3) If variables are independent then they are uncorrelated (r near zero), but the converse is not true
4) Sometimes, correlation value may mislead, as there may be some value of correlation by chance, but actually there is no evidence of correlation

5.2.4MERITS AND DEMERITS OF COEFFICIENT OF CORRELATION

Merits :

1) It is easy to understand and easy to calculate
2) It indicates type of correlation i.e. negative, positive or no correlation
3) It also gives clear information about extent of correlation, +1 for perfect positive and -1 for perfect negative correlation

Demerits :

1) It can mislead as higher correlation does not always mean close relationship. Two variables can have high value of correlation but may not actually have any relatinship
2) It is affected by extreme values of data set
3) Non linear relation is not very clearly indicated by correlation coefficient, whereas it is vlearly seen in Scatter plot

5.2.5 RANK CORRELATION

Rank correlation coefficient measures the degree of similarity between two rankings.

For example, in a singing competition, two judges may give their independent opinion about the participants through ranking, say 1, 2, 3 etc. With the Rank correlation coefficient, one can find the extent to which these two judges agree on the performance of the participant.

Spearman's Rank Correlation

Spearman's Rank Correlation, $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

Where d is difference in Rank

Ex 7 : Find Spearman's Rank Correlation

R1	R2
1	2
2	3
3	1
4	5
5	4

Solution :

R1	R2	d = R1 – R2	d^2
1	2	-1	1
2	3	-1	1
3	1	2	4
4	5	-1	1
5	4	1	1
		Total	8

$n = 5$

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{(6)(8)}{5(25-1)} = 1 - \frac{48}{120} = 1 - 0.4 = 0.6$$

$R = 0.6$ **(Ans)**

Spearman's Rank Correlation when Ranks are repeated

$$R = 1 - \frac{6\Sigma(d^2 + cf)}{n(n^2-1)}$$

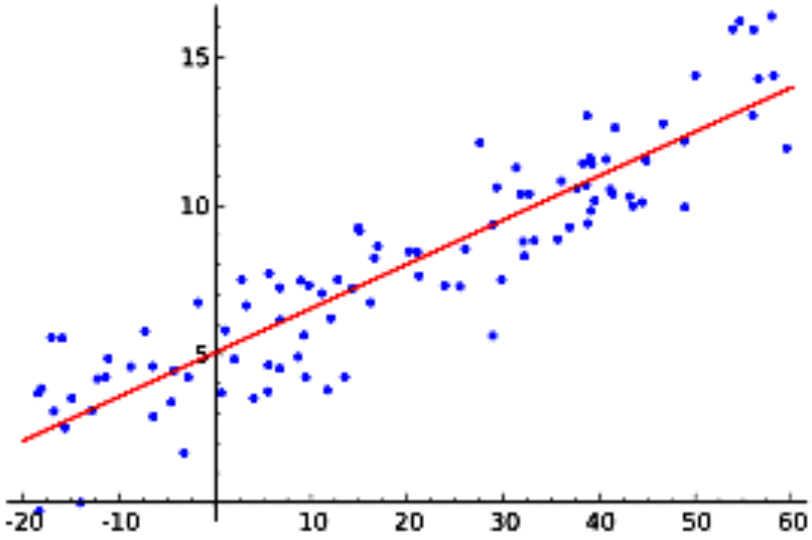
Where d is difference in Rank

$$cf = \frac{m(m^2-1)}{12}, \quad \textit{where } m \textit{ is the number of times Rank is repeated}$$

5.3 REGRESSION

Regression is an estimation technique. It uses historical data/information to estimate/predict near future value of that parameter. For Example, score of a student in Mathematics exam can be predicted based on student's performance in a few previous years.

Regression line :



If X is independent variable, and Y is dependent variable, then the Regression line can be given as :

$$Y = a + bX$$

Above Regression equation represents a strainght line. In practice, there can be non linear relationship between X and Y, in such a case, the Regression equation can include square or cube or higher degree terms also.

Regression Equation actually approximates and straightens the point orientation by introducing some error for alignment of the points to get a straight line .i.e. Regression line.

5.3.1 LINEAR REGRESSION USING METHOD OF LEAST SQUARES

Method of Least Squares is one of the methods to derive Regression Equation.

Two parameters $a \text{ and } b \text{ of the linear equation } Y = a + bX$ can be found out using two normal equations.

$\sum Y = an + b\sum X$ …….. Normal Equation I

$\sum XY = a\sum X + b\sum X^2$ …….. Normal Equation II

Solving these equations give values of **a** and **b** required to form Regression Equation

Ex 8 : Form Regression Equation for the following data set.

X	Y
5	12
12	15
15	32
22	35
25	45

Solution :

Regression Equation is $Y = a + bX$

The two Normal equations are :

$\sum Y = an + b\sum X$ …….. Normal Equation I

$\sum XY = a\sum X + b\sum X^2$ …….. Normal Equation II

	X	Y	XY	X^2
	5	12	60	25
	12	15	180	144
	15	32	480	225
	22	35	770	484
	25	45	1125	625
Total	**79**	**139**	**2615**	**1503**

Substituting these values in the two normal equations :

$139 = 5a + 79b$

$2615 = 79a + 1503b$

Solving simultaneously, or by method of substitution,

$a = 1.83$ *and* $b = 1.644$

Substituting these values in the Regression Equation :

$Y = 1.83 + 1.644X$ is the Regression Equation **(Ans)**

Ex 9 : Form Regression Equation for the following data set, and hence estimate

Y *for* $X = 10$

X	Y
1	25
3	18
4	12
6	5
9	1

Solution :

Regression Equation is $Y = a + bX$

The two Normal equations are :

$\sum Y = an + b\sum X$ …….. Normal Equation I

$\sum XY = a\sum X + b\sum X^2$ …….. Normal Equation II

	X	Y	XY	X^2
	1	25	60	25
	3	18	180	144
	4	12	480	225
	6	5	770	484
	9	1	1125	625
Total	**23**	**61**	**166**	**143**

Substituting these values in the two normal equations :

$61 = 5a + 23b$

$166 = 23a + 143b$

Solving simultaneously, or by method of substitution,

$a = 26.37 \quad and \quad b = -3.081$

Substituting these values in the Regression Equation :

$Y = 26.37 - 3.081X$ is the Regression Equation

For

$X = 10$, $Y = 26.37 - 3.081 * 10 = -4.44$

$Y = -4.44$ **(Ans)**

5.3.2 REGRESSION COEFFICIENT

Regression Coefficient b of Y on X

Regression Coefficient **b** of Y on X is given as :

$$b_{yx} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

Regression Equation can now be obtained as :

$Y - \bar{Y} = b_{yx}(X - \bar{X})$

Ex 10 : Find Regression Equation using Regression coefficient b_{yx}

X	Y
2	13
3	24
4	54
6	65
9	72

Solution :

	X	Y	XY	X^2
	2	13	60	25
	3	24	180	144
	4	54	480	225
	6	65	770	484
	9	72	1125	625
Total	**23**	**61**	**166**	**143**

Regression Coefficient **b** of Y on X is given as :

$$b_{yx} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

$$b_{yx} = \frac{(5)(166) - (23)(61)}{(5)(143) - (23)^2} = 8.364$$

$$b_{yx} = \mathbf{8.364}$$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{24}{5} = 4.8 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{228}{5} = 45.6$$

Regression Equation can now be obtained as :

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 45.6 = (8.364)(X - 4.8)$$

$\mathbf{Y = 5.45 + 8.364\,X}$ is the Regression Equation **(Ans)**

Regression Coefficient b of Y on X

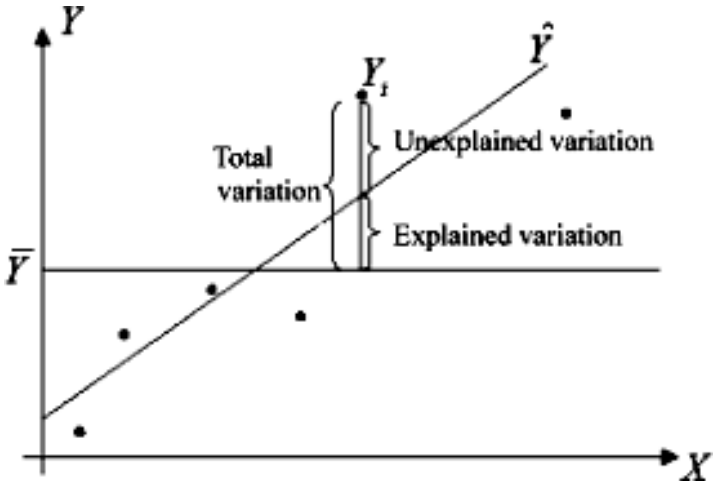
Regression Coefficient **b** of X on Y is given as :

$$b_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2}$$

Regression Equation can now be obtained as :

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

5.3.3 COEFFICIENT OF DETERMINATION



The Coefficient of detrmination, r^2, is a paramter used to judge how well the estimated Regression line fits all the data, where r, is Karl Pearson's coefficient of Correlation.

$$Coefficient\ of\ determination, \qquad r^2 = \frac{explained\ variation}{total\ variation}$$

If the Regression line passes through all or most points, then coefficient if determination will be close to 1.

Since, $-1 \leq r \leq 1, \qquad 0 \leq r^2 \leq 1$

Significance of coefficient of detrmination

1) It gives the strength of linear relaionship between two variables
2) It gives confidence to obtain variable to be predicted from the indepndent variable
3) The coefficient of determination is the ratio of explained variation to toal variation
4) It represents the quantum of data that is closest to the line of best fit
5) It is a measure of how well the Regression line represents the data

5.3.4 PROPERTIES OF REGRESSION COEFFICIENT

1) The point $(\bar{X}, \bar{Y})$ lies on both the Regression lines
2) In case of perfect correlation between two variables, $r = 1$ or $r = -1$
3) Slope of Regression equation Y on X is given as, b_{yx} wheras, slope of Regression equation X on Y is given as $\frac{1}{b_{xy}}$

4) The angle between two Regression lines is given as, $\theta = tan^{-1}\left|\frac{b_{yx}b_{xy}-1}{b_{xy}+b_{yx}}\right|$

5.4 SUMMARY

1) Correlation between two parameters can be represented either by Scatter Graph or Karl Pearson's coefficient of Correlation (r) can be used
2) Karl Perason's coefficient of correlation ranges between -1 to +1. Negative correlation has negative value of r and positiove correlation has positiove value of r
3) Regression line helps to estimate or predict **near future** value of the dependent parameter using historical values of the independent variable
4) Regression line can be found out using method of least squares or using Regression coefficient method
5) Coefficient of determination helps to understand how well is the regression line fits or covers all or most data points

5.5 EXERCISE

1) Plot Scatter Graph and comment

X	Y
201	34
226	45
230	56
312	53
340	62
357	64

2) Find Karl Pearson's coefficient of correlation

X	Y
55	12
43	10
32	7
24	4
18	3
11	1

3) Find Spearman's Rank Correlation

R1	R2
1	4
2	3
3	2
4	1
5	5

4) Find Regression Equation for the following data set, using method of least squares

X	Y
12	12
18	34
26	67
34	87
53	106
66	134

5) Find Regression Equation using Regression coefficient b_{xy}

X	Y
1	4
6	22
8	45
10	77
11	87

5.6 LIST OFREFERENCES

1) Probability, Statistics, design of experiments and queuing theory with applications of Compter Science, S. K. Trivedi, PHI

2) Applied Statistics, S C Gupta, S Chand

6

PROBABILITY

Unit Structure

6.0 OBJECTIVE

The study of Probability helps learner to find solution to various types problems which have some uncertainty in their occurence. Thie shapter explains various definitions, concept and terms used in probabiluty study in detail.

Learner should be able to understand and find solution to various problems for which probability theory gives reasonably good solution.

6.1 INTRODUCTION

Study of Probability is the study of chance. Probability theory is widely applied to understand economic, social as well business problems.

Refer to the statements used by us in our daily life :

1) The train may get delayed
2) There is a chance of getting distinction in Mathematics by Mahesh
3) Asha may come on time today

Such statements are commonly used by all of us. One can systematically study such probable events using principles of Probability discussed in this chapter.

6.2 SOME BASIC DEFINITIONS OF PROBABILITY

Experiment : An experiment is an action that has more than one posiible outputs

For Example :

1) Tossing a coin gives either a Head or a Tail
2) Throwing a die gives any one number from 1 to 6 on top face of the die
3) A student appearing for an exam may pass or may fail exam

Experiment may be random or deterministic.

The output of the random experiment changes and occurs randomly without any bias. In random experiments, all outcomes are equally likely. For example, tossing a coin

The outcome of the deterministic experiment does not change when performed many times. For example, counting number of windows of a particular room

Outcome : The result of an experient is called outcome. For example, counting number of students in a class

Trial : Performing an experiment is called taking a trial

Sample Space : The collection of all possible outcomes is called sample space of that experiment, For example, drawing a ball from a box having three balls of Red, Blue and Green colours has a sample space of balls of Red, Blue and Green colours. Sample space is demoted by letter S

Sample point : Each outcome of the sample space is called sample point. The total number of sample points are denoted as n(S)

Finite sample space : When the number of outcomes are finite, the sample space in finite sample space. For example, number of students in Statistics class of a college

Countably infinite sample space : When the number of elements in a sample space are infinite, the sample space is said to countably infinite sample space. For example, set all all natural numbers

Exhaustive outcomes : Outcomes are exhaustive if they combine to be the entire sample space.For example, outcomes Head and Tail are exhaustive outcomes, when a coin is tossed

Event : Any subset of sample space associated with random experiment is called an Event. Fro example, for a sample space={1, 2, 3, 4, 5}, an event A can be “getting and odd number” and can be written as A={1, 3, 5}

Types of Event : Events can be described as given below :

1) Simple event : An event having only one outcome is called simple event. For example, the evet of getting a head when a coin is tossed
2) Impossible event : The event corresponding to null set is called an impossible event. For example, an event of getting a number more than 6 when a die is thrawn
3) Sure event : The event corresponding to the sample space is called sure event. For example, an event of getting either a head or a tail when a fair coin is tossed
4) Mutually exclusive events : Two or more events are said be mutually exclusive events if they do not have a sample point in common. For example, an event of getting an even and another event of getting an even number when a die is rolled
5) Exhaustive events : The events are said to be exhaustive events if occurrence of any one event is surely going to take place. For example, event of getting either red or black card when a card is drawn from a pack of cards
6) Equally likely event : When all events have same chance of occurrence then the events are equally likey. For example, getting a Head or a Tail when an uniased coin is tossed, are called equally likely events
7) Independent events : Two or more events are said to be independent events if one of them is not affected by occurrence of any other events. i.e. P(A/B)=P(A)

6.3 PERMUTATIONS AND COMBINATIONS

Factorial: Factorial of a real number n is written as $n!$ such that
$n! = n.(n-1).(n-2).....2.1$

Ex 1 : Find $5!$
Solution : $5! = 5.4.3.2.1 = 120$ **(Ans)**

Permutation : Permutation means arrangement of objects in different ways. For example, out of three objects A, B and C taken two at a time can be arranged as AB, BA, BC, CB, CA, AC. We can arrange in six different ways, as order or sequence of objects in Permutations is important. So, if n objects are are arranged taken r at a time can be written as,

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

Ex 2 : Find $^4P_3$
Solution : take $n = 4 \text{ and } r = 3$

$$^nP_r = \frac{n!}{(n-r)!} = \frac{4!}{(4-3)!} = \frac{4!}{1!} = \frac{4.3.2.1}{1} = 24$$

$^4P_3 = 24$ **(Ans)**

Ex 3 : How many ways are there for eight men and five women to stand in a row so that no two women stand next to each other.

Solution :

Eight men can be arranged in $8! = 40320$ ways.

Five women can be arranged in 9 ways as shown below :

* M * M * M *M *M *M *M *M *

Here * represents a place for a woman, and M represents a place for man.

Five women can be arranged in 9 places in

$$^9P_5 = \frac{9!}{(9-5)!} = \frac{9!}{(9-5)!} = \frac{9!}{4!} = \frac{9.8.7.6.5.4!}{4!} = 15120$$ ways.

So, together eight men **and** five women can be arranged such that no two women stand together as :

Total number of ways = $8! \,.\, {}^9P_5 = 40320 \,.15120 = 60,96,38,400$ ways **(Ans)**

Ex 4 : In Hhw many ways can the letters of the word 'MOUSE' arranged, where meaning/spelling does not matter.

Solution :
The words can be arranged in $^5P_5 = 5! = 120$ ways. **(Ans)**

<u>Combination</u> : Combination is a selection of objects without consideting the order of arrangements. For Example, for three objects A, B and C, when two objects are taken at a time, the arrangement can be AB, BC and AC. Order or sequence of arranements is not important in case of Combination. So, Combination of n objects taken r at a time can be written as,

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

Ex 5 : Find $^4C_3$

Solution : take $n = 4 \text{ and } r = 3$

$$^nC_r = \frac{n!}{r!\ (n-r)!} = \frac{4!}{3!\ (4-3)!} = \frac{4!}{3!\ 1!} = \frac{4.3.2.1}{3.2.1.1} = 4$$

$^4C_3 = 4$ **(Ans)**

Ex 6 : Find $^3P_2 + {}^4C_2$

Solution :

$$^{n}P_{r} = \frac{n!}{(n \quad r)!} = \frac{3!}{(3 \quad 2)!} = \frac{3!}{1!} = \frac{3.2.1}{1} = 6$$

$$^{3}P_{2} = 6$$

Also,

$$^{n}C_{r} - \frac{n!}{r!\ (n-r)!} - \frac{4!}{2!\ (4-2)!} - \frac{4!}{2!\ 2!} - \frac{4.3.2.1}{2.1.2.1} - 6$$

$$^{4}C_{2} = 6$$

$$\therefore\ ^{3}P_{2} + {}^{3}P_{2} = 6 + 6 = 12$$ **(Ans)**

Ex 7 : In how many ways can a committee of 2 officers and 3 clerks can be made from 4 officers and 10 clerks.

Solution : This can be done in $^{4}C_{2} * {}^{10}C_{3}$ ways

$$^{4}C_{2} = \frac{4!}{2!\ (4-2)!} = \frac{4!}{2!\ 2!} = \frac{4.3.2.1}{2.1.2.1} = 6$$

$$^{10}C_{3} = \frac{10!}{3!\ (10-3)!} = \frac{10!}{3!\ 7!} = \frac{10.9.8.7!}{3.2.1\ .7!} = 120$$

$$^{4}C_{2} * {}^{10}C_{3} = 6 * 120 = 720$$ ways **(Ans)**

6.4 CLASSICAL AND AXIOMATIC DEFINITIONS OF PROBABILITY

Classical definition of Proability

When a random experiment is conducted having sample space S having n(S) equally likely outcomes, the event A having n(A) favourable outcomes, the probability of occurrence of event A is given as P(A) such that :

$$P(A) = \frac{n(A)}{n(S)}$$

Some inportant points regarding Probability definition are :

1) The sum of all probabilities in the sample space is 1 (one)
2) The probability of an impossible event is 0 (zero)
3) The probability of a sure event is 1 (one)
4) The probability of not occuring an event is 1 – probability of occuring an event. i.e. $P(\bar{A}) = 1 - P(A)$

Ex 8 : Write down sample space for each of the following cases

1) A coin is thrawn three times

2) A coin is thrawn three times and number of heads in each thraw is noted

3) A tetraheadron (a solid with four traingular surfaces) whose sides are painted red, red, blue and green. The color of the side touching the gound is noted

4) Blood group of husband and wife are tested and noted

Solution

1) $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

2) $S = \{0, 1, 2, 3\}$

3) $S - \{red, blue, green\}$

4) $S = \{(O,O),(O,A),(O,B),(O,AB),(A,O),(A,A),(A,B),(A,B),(A,AB),$
$(B,O),(B,A),(B,AB),(AB,O),(AB,A),(AB,B),(AB,AB)\}$

Ex 9 : Thre eunbiased coins are tossed. What is the probability of getting at least one Head.

Solution :

Sample Space, $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

$n(S) = 8$

Let A be the event of getting at least one Head

$A = \{HHH, HHT, HTH, HTT, THH, THT, TTH\}$

$n(A) = 7$

$$P(A) = \frac{n(A)}{n(S)} = \frac{7}{8}$$ **(Ans)**

Ex 10 : Nine tickets are marked numbers 1 to 9. One ticket is drawn at random. What is the probability that the number is an odd number.

Solution :

$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

$n(S) = 9$

$A = \{1, 3, 5, 7, 9\}$

$n(A) = 5$

$$P(A) = \frac{n(A)}{n(S)} = \frac{5}{9}$$ **(Ans)**

Ex 11 : An urn contains 8 blue balls, 7 green balls and 5 red balls. One ball is drawn at random, what is the probability that it is (a) a red ball, (b) a blue ball.

Solution :

$n(S) = 8 + 7 + 5 = 20$

(a) Let A be the event of getting a red ball

$n(A) - 5 \quad P(A) = \frac{n(A)}{n(S)} = \frac{5}{20} = \frac{1}{4}$

(b) Let B be the event of getting a blue ball

$n(B) = 8 \quad P(B) - \frac{n(B)}{n(S)} - \frac{8}{20} - \frac{2}{5}$

Ex 12 : From a well shuffeled pack of cards, a card is drawn at random. What is the probability that the drawn card is a red card

Solution :

$n(S) = 52$

Let A be the event of getting a red card

$n(A) = 26 \quad P(A) = \frac{n(A)}{n(S)} = \frac{26}{52} = \frac{1}{2}$

$P(A) = \frac{1}{2}$ **(Ans)**

Ex 13 : What is the probability of getting a sum nine (9) when two dice are thrawn

Solution :

$n(S) = \{(1,1), (1,2), \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, (5,6), (6,6)\} = 36$

Let A be the event of getting a sum nine (9)

$A = \{(3,6), (4,5), (5,4), (6,3)\}$

$n(A) = 4$

$P(A) = \frac{n(A)}{n(S)} = \frac{4}{36} = \frac{1}{9}$ **(Ans)**

Ex 14 : The Board of Directors of a company wants to form a quality management committee to monitor quality of their products. The company has 5 scientists, 4 engineers and 6 accountants. Find the probability that the committee will have 2 scientists, 1 engineer and 2 accountants.

Solution :

$n(S) = {}^{15}C_5 = \frac{15!}{5!\ (15-5)!} = \frac{15!}{5!\ 10!} = \frac{15.14.13.12.11.10!}{5.4.3.2.1.10!} = 3003$

Let A be the event of having 2 scientists, 1 engineer and 2 accountants

$n(A) = {}^5C_2 . {}^4C_1 . {}^6C_2 = 10.4.15 = 600$

$P(A) = \frac{n(A)}{n(S)} = \frac{600}{3003}$

$P(A) = \frac{600}{3003}$ **(Ans)**

Axiomatic definition of Proability

Suppose, for an experiment, S is the sample space containing outcomes, $S_1, S_2, S_3 \ldots\ldots, S_n$, then assigning a real number $P(S_i)$ to each $S_i \in S$ uch that

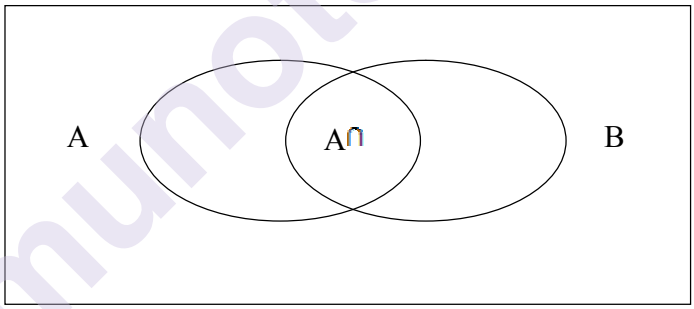
1) $0 < P(S_i) < 1$
2) $\sum P(S_i) = 1$

6.5 ADDITION THEOREM

If A and B are two events defined on sample space, S then

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

a) Addition theorem can also be explained by Venn diagram



b) If two events are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$

c) For three events,

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + -P(A \cap B \cap C)$

Ex 15 : An integer is chosen at random from 1 to 100. Find the probability that it is multiple of 5 or a perfect square

Solution :

$n(S) = 100$

Let A be the event of getting a number multiple of 5

Let B be the event of getting a perfect square

$A = \{5, 10, 15, \ldots\ldots, 95, 100\}$

$n(A) = 20$

$$P(A) = \frac{n(A)}{n(S)} = \frac{20}{100}$$

$B = \{1, 4, 9, \ldots..81, 100\}$

$n(B) = 10$

$$P(B) = \frac{n(B)}{n(S)} = \frac{10}{100}$$

$A \cap B = \{25, 100\}$

$n(A \cap B) = 2$

$$P(A \cap B) - \frac{n(A \cap B)}{n(S)} - \frac{2}{100}$$

By addition theorem,

$P(A \cup B) = P(A) + P(B) \quad P(A \cap B)$

Required probability of getting a multiple of 5 or a getting a perfect square is $P(A \cup B)$

$$P(A \cup B) = \frac{20}{100} + \frac{10}{100} - \frac{2}{100} = \frac{28}{100} = 0.28$$

$P(A \cup B) = 0.28$ **(Ans)**

Ex 16 : A card is drawn at random from a pack of cards. Find the probability that the drawn card is a diamond or face card.

Solution :

$n(S) = 52$

Let A be the event of getting a diamond card

Let B be the event of getting a face card

$n(A) - 13$

$$P(A) = \frac{n(A)}{n(S)} = \frac{13}{52}$$

$n(B) = 12$

$$P(B) = \frac{n(B)}{n(S)} = \frac{12}{52}$$

$A \cap B = \{a\ face\ card\ of\ diamond\}$

$n(A \cap B) = 3$

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{3}{52}$$

By addition theorem,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Required probability of getting a multiple of 5 or a getting a perfect square is $P(A \cup B)$

$$P(A \cup B) = \frac{13}{52} + \frac{12}{52} \quad \frac{3}{52} - \frac{22}{52} - \frac{11}{26}$$

$$P(A \cup B) = \frac{11}{26}$$ **(Ans)**

6.6 CONDITIONAL PROBABILITY

Let there be two events A and B. The probability of event A given that event B has occurred is known as conditional probabilityof A given that B has occurred and is given as :

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$$

Ex 17: Given $P(A) = \frac{2}{3}$, $P(B) = \frac{1}{4}$ *and* $P(A \cap B) = \frac{1}{5}$. Find $P\left(\frac{A}{B}\right)$

Solution :

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{5}}{\frac{1}{4}} = \frac{4}{5}$$ **(Ans)**

Ex 18 : Find the probability that a single toss of die will result in a number less than 4 if it is given that the toss resulted in an odd number.

Solution : Let event A be the toss resulting in an odd number

And let event B be getting the number less than 4

$A = \{1, 3, 5\}$

$n(A) = 3$

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{6}$$

$B = \{1, 2, 3\}$

$n(B) = 3$

$A \cap B = \{1, 3\}$

$n(A \cap B) = 2$

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{2}{6}$$

$$P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}$$ **(Ans)**

6.7 BAYE'S THEOREM

Let $A_1, A_2 \ldots \ldots A_n$ be a set of mutually exclusive events that together form the sample space S. Let B be any event from the same sample space. Then Baye's theorem states that

$$P\left(\frac{A_i}{B}\right) = \frac{P(A_i).P\left(\frac{B}{A_i}\right)}{P(A_1).P\left(\frac{B}{A_1}\right) + P(A_2).P\left(\frac{B}{A_2}\right) + \ldots + P(A_n).P\left(\frac{B}{A_n}\right)}$$

Ex 19 : In a toy factory, machines A_1, A_2 *and* A_3 manufacture respectively 25%, 35% and 40% of total toys. Of these 5%, 4% and 2% are defective toys. A toy is selected at random and is found to be deefctive. What is the probability that it was manufactured by machine A_2

Solution :

$P(A_1) = P(\text{that the machine } A_1 \text{ manufacture toys}) = 25\% = 0.25$

$P(A_2) = P(\text{that the machine } A_2 \text{ manufacture toys}) = 35\% = 0.35$

$P(A_3) = P(\text{that the machine } A_3 \text{ manufacture toys}) = 40\% = 0.40$

Let B be any event that the drawn toy is defective.

$$P\left(\frac{B}{A_1}\right) = P(\text{the defective toy was from machine } A_1) = 5\% = 0.05$$

$$P\left(\frac{B}{A_2}\right) = P(\text{the defective toy was from machine } A_2) = 4\% = 0.04$$

$$P\left(\frac{B}{A_3}\right) = P(\text{the defective toy was from machine } A_1) = 2\% = 0.02$$

We have to find $P\left(\frac{A_2}{B}\right)$

$$P\left(\frac{A_2}{B}\right) = \frac{P(A_2).P\left(\frac{B}{A_2}\right)}{P(A_1).P\left(\frac{B}{A_1}\right) + P(A_2).P\left(\frac{B}{A_2}\right) + P(A_3).P\left(\frac{B}{A_3}\right)}$$

$$P\left(\frac{A_2}{B}\right) = \frac{(0.35)(0.04)}{(0.25)(0.05) + (0.35)(0.04) + (0.40)(0.02)} = 0.40$$

Required probability is **0.40** **(Ans)**

6.8 SUMMARY

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

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

3) $P(A) = \frac{n(A)}{n(S)}$

4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

5) $P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$

6) $$P\left(\frac{A_i}{B}\right) = \frac{P(A_i).P\left(\frac{B}{A_i}\right)}{P(A_1).P\left(\frac{B}{A_1}\right) + P(A_2).P\left(\frac{B}{A_2}\right) + \ldots + P(A_n).P\left(\frac{B}{A_n}\right)}$$

6.9 EXERCISE

1) One card is drwan at random from a pack of cards. What is the probability that it is a King or a Queen.

2) Find $^4C_1$

3) Given an equiprobable sample space $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$, and an event $A = \{0, 4, 7\}$ Fnd $P(A)$ *and* $P(A^-$)

4) Given, $P(A) = 0.7,$ $P(B) = 0.5$ *and* $P(A \cap B) = 0.3$ Find $P(A \cup B)$

5) A class has 40 boys and 20 girls. How many ways a class representative (CR) be selected such that the CR is either a boy or a girl

6) From a set of 16 tickets numbered from 1 to 16, one ticket is drawn at random. Find the probability that the number is divisible by 2 or 5

7) A car manufacturing company has two plants. Plant A manufactures 70% of the cars and the plant B manufactures 30 % of the cars. 80% and 90% of the cars are of standard quality at plant A and plant B respectuvely. A car is selected at random and is found to be of standard quality. What is the probability that is was manufactured in plant A

6.10 LIST OFREFERENCES

1) Probability, Statistics, design of experiments and queuing theory with applications of Compter Science, S. K. Trivedi, PHI
2) Applied Statistics, S C Gupta, S Chand

❖❖❖❖