Tulsiramji Gaikwad-Patil College of Engineering & Technology

Department of Master of Computer Application

Subject Notes

Academic Session: 2018 – 2019

Subject: Statistics & Numerical Mathematics

Semester: II

UNIT – I

1. **(A) Define Statistics and Explain the scope of statistics in various disciplines in detail.**

Ans: Defination of Statistics: Statistics has been defined differently by different authors from time to time. The reasons for a variety of definitions are primarily two. *First,* in modem times the fIeld of utility of Statistics has widened considerably. In ancient times Statistics was confined only to the affairs of State but now it embraces almost every sphere of human activity. Hence a number of old definitions which'were confined to a very narrow field of eQquiry, were replaced by new definitions which are much more cOl1)prehensive and exhaustive. *Secondly,* Statistics has been defined in two ways. Some writers define it as' *statistical data', i.e.,* numerical statement of facts,while others define itas *'statistical methods', i.e.,* complete body of the principles and techniques used inrcollecting and analysing suc.h data. Some of the i~portant definitions are given below.

Statistics as 'Statistical Data'

Webster defines Statistics ali "classified facts represt;nting the conditions of the people in a State ... especially those facts which can be stated in numbers or in any other tabular or classified arrangement." This definition, since it confines Statistics only to the data pertaining to State; is inadequate as the domain of .Statistics is much wider.

Bowley defines Statistics as " numerical statements of facts in any department of enquiry placed in· relation to each other." A more exhaustive definition is given by Prof. Horace Secrist as follows:

" By Statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable

standards of accuracy, collected in a systematic manner for a pre"lletermined purpose and placed in relation tv each other."

Statistics as Statistical,Methods

Bowley himselr d.efines Statistics in. the rollQwing three different ways:

(i) Statistics may be called the science of counting.

(ii) Statistics may rightly be called the science of.averages.

(iii) Statistics is the science of the measurement of social organism, regarded all a whole in all its manifestations.

But none of the above definitions is adequate. The *first* because1tatisticsJis not merely confined to the collection of data as other aspects like presentation, analysis and interpretation, etc., are also covered by it. The *second*, because averages are onl y a part ofthe statistical tools used in the analysis of the data, others' being dispersion, skewness, kurtosi'S, correlation, regression, etc. J'he *third*, because it restricts the application of Statistics to sociology alone while in modem days Statistics is used in almost all sciences - social as well as physical. According to Boddington, " *Statistits is the.. science of estimates and probabilities."* This also is an inadequate definition smce probabilities'and estimates] constitute only a part of the statistical methods.

scope of statistics

In modern times, statistics is viewed not as a mere device for collecting numerical data but as a means of developing sound techniques for their handling and analysis and drawing valid inferences from them. As such it is not confined to the affairs of the state but is intruding constantly into various diversified spheres of life – social as well as physical – such as biology, psychology, education, economics business management etc.

We now discuss briefly the importance of statistics in some different sectors and disciplines.

1. **Statistics and Planning:** Statistics is indispensable to planning. In the modern age which is termed as “the age of planning”, almost all organizations in the government or managements of business are resorting to planning for efficient working and for formulating policy decision. To achieve this end, the statistical data relating to production, consumption, prices, investment, income, expenditure, etc and various advance statistical techniques for handling and analyzing such complex data are of paramount importance.

2. **Statistics and Mathematics:** Statistics is intimately related to and essentially dependent upon mathematics. The modern theory of statistics has its foundations on the theory of probability which in turn is a particular branch of more advanced mathematical theory of measure and Integration. The main stalwarts in the theory of modern statistics namely Laplace, Bernoulli, Pascal, De-Moivre. Even increasing role of mathematics in statistics has resulted in the development of a new branch of Statistics called "*Mathematical Statistics*".
3. **Statistics and Economics:** Statistical data and techniques of statistical analysis have proved immensely useful in solving a variety of economic problem, such as wages, process, consumption, production, distribution of income and wealth etc. Statistical tools like's Index numbers, Time series Analysis, Demand Analysis and Forecasting Technique are extensively used for efficient planning and economics development of a country.
4. **Statistics and Business:** Statistics is an indispensable tool of production control also. Business executives are relying more and more on statistical techniques for studying the needs and the desires of the consumers and for many other purposes. The success of businessman more or less depends upon the accuracy and precision of his statistical forecasting. Wrong expectation, which may be the result of faulty and inaccurate analysis of various causes affecting a particular phenomenon, might lead to its disaster.

Statistical techniques have also been used widely by business organization in:

(i) Carrying out Time and Motion studies.

(ii) Investment (based on the statistical analysis of consumer preference studies – demand analysis).

(iii) Personal Administration (for the study of statistical data relating to wages, cost of living, incentive plans, effect of labor dispute/ unrest on the production, performance standards etc).

(iv) Credit Policy

(v) Inventory Control

(vi) Dale control

5. **Statistical and Biology, Astronomy and Medical Science:** The association between statistical methods and biological theories was first studies by Francis Galton in his work in Regression.

(B) Derive median formula for continuous frequency distribution and hence find median for the following distribution: (S-14)

X	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
F	10	25	30	9	8	3

Ans: Derivation of the Median Foemula

Let us consider the following continuous frequency distribution, ($x_1 < x_2 < \ldots\ldots < x_{n+1}$):

Class interval: $x_1 - x_2$ $x_2 - x_3$ ..... $x_k - x_{k+1}$ ..... $x_n - x_{n+1}$

Frequency : f_1 f_2 ..... f_k ..... f_n

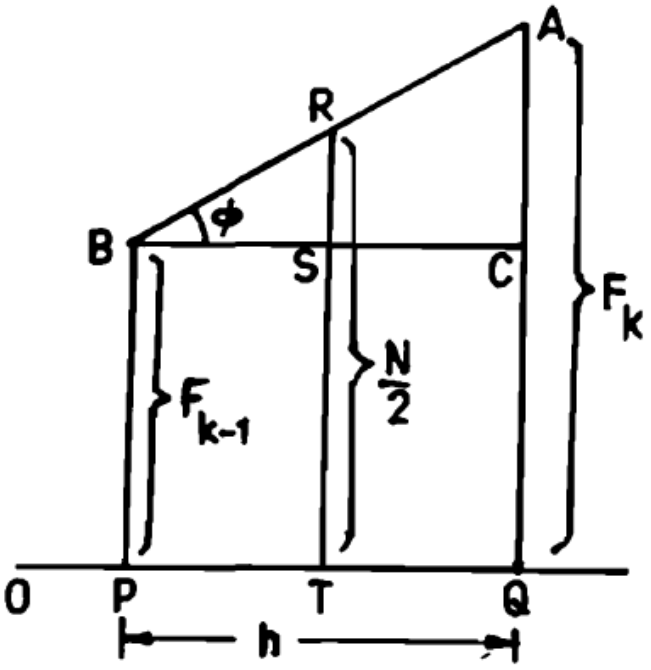
The cumulative frequency distribution is given by:

Class interval: $x_1 - x_2$ $x_2 - x_3$ ..... $x_k - x_{k+1}$ ..... $x_n - x_{n+1}$

Frequency: F_1 F_2 ..... F_k ..... F_n

where $F_i = f_1+f_2+ \ldots\ldots\ldots+f_i$. The class $x_k - x_{k+1}$ is the median class if and only if $F_{k-1} < 1/2N < F_k$

Now, if we assume that the variate values are uniform distributed over the median class which implies that the ogive is a straight line in the median class, then we ger form the fig.



$$\tan \emptyset = \frac{RS}{BS} = \frac{AC}{BC}$$

i.e. $$\frac{RT-TS}{BS} = \frac{AQ-CQ}{BC}$$

or $$\frac{RT-BP}{BS} = \frac{AQ-BP}{PQ}$$

or $$\frac{\frac{1}{2}N - F_{k-1}}{BS} = \frac{F_k - F_{k-1}}{PQ} = \frac{F_k}{h}$$

where f_k is the frequency and h the magnitude of the median class.

$$BS = \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)$$

Hence

Median – OT = OP + PT = OP + BS = l + $\frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)$, which is the required formula.

X	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
F	10	25	30	9	8	3

Computation of Median

x	f	c.f.
0 – 10	10	10
10 – 20	25	35
20 – 30	30	65
30 – 40	9	74
40 – 50	8	82
50 – 60	3	85

Total	N = 85	

Here 1/2N = 1/2(85) = 42.5.

Cumulative frequency just greater than 42.5 is 65 and the corresponding class is 20 – 30.

Thus median class is 20 – 30 .

Hence using median formula

Median = $l + \frac{h}{f}(\frac{N}{2} - C)$

Where l is lower limit of the median class,

f is the frequency of the median class,

f is the magnitude of the median class,

'c' is the c.f. of the class preceding the median class,

Median = $20 + \frac{10}{30}(\frac{85}{2} - 35)$

$= 20 + \frac{10}{30}(42.5 - 35)$

$= 20 + \frac{10}{30}(7.5) = 20 + \frac{75}{30} = 20 + 2.5 =$ **22.5**

(c) The geometric mean of 10 observation on a certain variable was calculated as 11.9. It was discovered that one of the observation was wrongly recorded as 13.3; in fact it was 31.3. Apply appropriate correction and calculate the correct geometric mean. (S-14)

Ans:

The geometric mean G of n observations x1,x2, ... xn is given by:

$G = (x_1 . x_2 \ldots . x_n)^{\frac{1}{n}}$

If x_1 is the observation copied wrongly instead of correct value x_1', then the corrected geometric mean G' is given by:

$G' = (x_1' . x_2 \ldots . x_n)^{\frac{1}{n}}$

$= (\frac{x_1'}{x_1} . x_2 \ldots . x_n)^{\frac{1}{n}}$

$= (x_1 . x_2 \ldots\ldots x_n)^{\frac{1}{n}} (\frac{x_1'}{x_1})^{\frac{1}{n}}$

$= G . (\frac{x_1'}{x_1})^{\frac{1}{n}}$

In the given problem, G= 11.9 , n = 10 , x_1 = 13.3 , x_1' = 31.3

$$\therefore Corrected\ G.G.(G') = 11.9 \times (\frac{31.3}{13.3})^{\frac{1}{10}}$$

$$\log_{10} G' = \log_{10} 11.9 + \frac{1}{10}(\log_{10} 31.3 - \log_{10} 13.3)$$

$$= 1.075 + \frac{1}{10}(1.49 - 1.12)$$

$$= 1.075 + 0.037$$

$$\log_{10} G' = 1.112$$

$$G' = Antilog\ (1.112)$$

G' = 12.94

(D) Find the missing frequencies of the 50 observations where the distribution as follows:

X	**0 – 10**	**10 – 20**	**20 – 30**	**30 – 40**	**40 – 50**
F	**?**	**12**	**13**	**?**	**4**

The mean of distribution is 22. (S-14)

Ans:

Lets we consider that missing frequencies are f1 and f2.

x	f	Mid Point (x)	fx
0 – 10	f1	5	5f1
10 – 20	12	15	180
20 – 30	13	25	325

30 – 40	f2	35	35f2
40 – 50	4	45	180
Total	50		

f1 + 12 + 13 + f2 + 4 = 50

29 + f1+ f2 = 50

f1 + f2 = 50 – 29

f1 + f2 = 21

f1 = 21 – f2 --------- (1)

Arithmetic mean

$$\bar{x} = \frac{1}{N}\sum_{i=0}^{n} fx$$

$$22 = \frac{1}{50}(5f1 + 180 + 325 + 35f2 + 180)$$

$$22 = \frac{1}{50}(5f1 + 35f2 + 685)$$

1100 = 5f1 + 35f2 + 685)

415 = 5f1 + 35f2 ------- (2)

Put the value of (1) in equation (2), we get

415 = 5 (21 – f2) + 35f2

415 = 105 – 5f2 + 35f2

415 – 105 = 30f2

310 = 30f2

$$\therefore f2 = \frac{310}{30}$$

f2 = 10.33 ≅ 10

Put the value of f2 in equation (1)

f1 = 21 – 10 = 11

(e) The geometric mean of 10 observations on a certain variable was calculated as 16.2. It was discovered that one of the observations was wrongly recorded as 12.9, in fact it was 21.9. Apply appropriate correction and calculate the correct geometric mean. (W – 14)

Ans: -

The geometric mean G of n observations x1,x2, ... xn is given by:

$G = (x_1.x_2 \ldots\ldots x_n)^{\frac{1}{n}}$

If x_1 is the observation copied wrongly instead of correct value x_1', then the corrected geometric mean G' is given by:

$G' = (x_1'.x_2 \ldots\ldots x_n)^{\frac{1}{n}}$

$= (\frac{x_1'}{x_1}.x_2 \ldots\ldots x_n)^{\frac{1}{n}}$

$= (x_1.x_2 \ldots\ldots x_n)^{\frac{1}{n}} (\frac{x_1'}{x_1})^{\frac{1}{n}}$

$= G \,.\, (\frac{x_1'}{x_1})^{\frac{1}{n}}$

In the given problem, G= 16.2, n = 10 , $x_1 = 12.9$, $x_1' = 21.9$

$$\therefore Corrected\ G.G.(G') = 16.2 \times (\frac{21.9}{12.9})^{\frac{1}{10}}$$

$$\log_{10} G' = \log_{10} 16.2 + \frac{1}{10}(\log_{10} 21.9 - \log_{10} 12.9)$$

$$= 1.2095 + \frac{1}{10}(1.3404 - 1.1106)$$

$$\log_{10} G' = 1.23248$$

$$G' = Antilog\ (1.23248)$$

G' = 17.08

(f) Explain origin and development of statistics and also give limitations of it. (W – 14)

Ans: Statistics, in a sense, is as old as the human society itself. Its origin can be traced to the old days when it 'was regarded as the 'science of State-craft' and was the by-

product of the administrative activity of the State. The word 'Statistics' seems to have been derived from the Latin word' status' or the Italian word' statista' or the German word' statistik' each of which means a 'political state'. In ancient times, the government used to collect .the information regarding the population and 'property or wealth' of the country-the former enabling the government to have an idea of the manpower of the country (to safeguard itself against external aggression, if any), and the latter providing it a basis for introducing news taxes and levies'.

In India, an efficient system of collecting official and administrative statistics existed even more than 2,000 years ago, in particular, during the reign of Chandra Gupta Maurya (324 -300 B.C.). From Kautilya's Arthshastra it is known that even before 300 B.C. a very good system of collecting 'Vital Statistics' and registration of births and deaths was in vogue. During Akbar's reign (1556 – 1605 A.D.), Raja Todarmal, the then land and revenue minister, maintained good records of land and agricultural statistics. In Aina,e-Akbari written by Abul Fazl (in 1596 - 97), one of the nine gems of Akbar, we find detailed accounts of the administrative and statistical surveys conducted during Akbar's reign.

Modem veterans in the development of the subject are Englishmen. Francis Gilton (1822-1921) , with his works on 'regression' , pioneered the use of statistical methods in the fiel(J of Biometry. Karl Pearson (1857-1936), the founder of the greatest statistical laooratory in England (1911), is the pioneer in. Correlation analysis. His discovery of the 'chi square test', the first and the most important of modem tests of significance, won for Statistics a place as a science, In 1908 the discovery of Student's 't' distribution by W.S. Gosset who' wrote under the. pseudonym of 'Student' ushered in an era of exact sample tests (small samples).

(g) Derive the modal formula for continuous frequency distribution and hence find mode for the following distribution: (W – 14)

x	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
f	25	22	21	15	26	30

Ans: Derivation of the mode formula:

Let us consider the continuous frequency distribution:

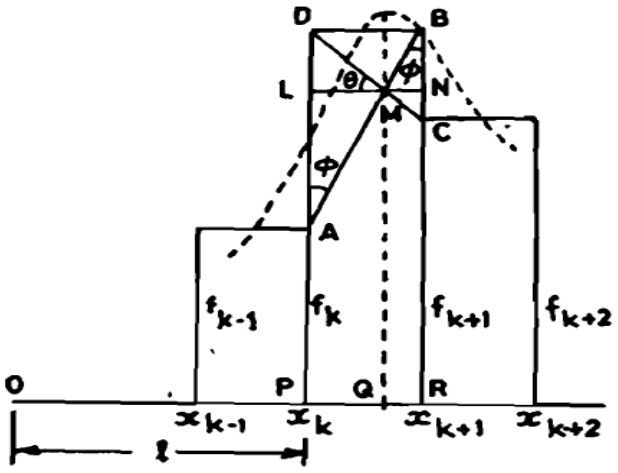
Class : $x_1 - x_2$ $x_2 - x_3$ ... $x_k - x_{k+1}$ ... $x_n - x_{n+1}$

Frequency: f_1 f_2 ... f_k ... f_n

Where all the classes are equal magnitude, say, h units.

If f_k is the maximum of all the frequencies, then the modal calss is $(x_k - x_{k+1})$.

Let us further consider a portion of the histogram, namely, the rectangles erected on the modal class and the two adjacent classes. The mode is the value of x for which the frequency curve has maxima. Let the modal point Q.



For the adjoining figure,

$$\tan\theta = \frac{LD}{LM} = \frac{NC}{MN}$$

And

$$\tan\phi = \frac{LM}{AL} = \frac{MN}{NB}$$

$$\frac{LM}{MN} = \frac{LD}{NC} = \frac{AL}{NB} = \frac{AL + LD}{NB + NC} = \frac{AD}{BC}$$

$$i.e. \quad \frac{LM}{LN - LM} = \frac{PD - AP}{BR - CR}$$

$$or \quad \frac{LM}{LN - LM} = \frac{f_k - f_{k-1}}{f_k - f_{k+1}}$$

Where h is the magnitude of the modal class. Thus solving for LM, we get

$$LM = \frac{h\,(f_k - f_{k-1})}{(f_k - f_{k+1}) + (f_k - f_{k-1})} = \frac{h\,(f_k - f_{k-1})}{2f_k - f_{k-1} - f_{k+1}}$$

Hence Mode = OQ = OP + PQ = OP + LM = l+ $\frac{h\,(f_k - f_{k-1})}{2f_k - f_{k-1} - f_{k+1}}$

x	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
f	25	22	21	15	26	30

Here maximum frequency is 30. Thus the class 50 – 60 is the modal class. Using mode formula, the value of mode is given by:

Mode = l+ $\frac{h\,(f_1 - f_0)}{2f_1 - f_0 - f_2}$

$$= 50 + \frac{10\,(30-26)}{(2\times30-26-0)}$$

$$= \qquad 50 + \frac{40}{34} \qquad = \qquad 50 + \frac{40}{34} = 51.176$$

(h) Derive the Geometric mean formula of the combined group and give merits and demerits of Geometric mean. (W – 14)

Ans: Geometric mean formula of the combined group :

If n1 and n2 are the sizes, G1 and G2 the geometric mean of two series respectively, the geometric mean G, of the combined series is given by:

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

Proof: Let x_{1i} (i=1,2,...,n_1) and x_{2j} (j = 1,2,...,n_2) be n1 and n2 items of two series respectively. Then by def.,

$G_1 = (x_{11} \cdot x_{12} \cdot \ldots x_{1n1})^{1/n1}$ => $\log G_1 = \frac{1}{n_1} \sum_{i=1}^{n1} \log x_{1i}$

$$G_2 = \left(x_{21} . x_{22} \ldots x_{2n_2}\right)^{1/n2} \Rightarrow \log G_2 = \frac{1}{n_2} \sum_{j=1}^{n2} \log x_{2j}$$

The geometric mean G of the combined series is given by:

$G = (x_{11} \cdot x_{12} \ldots x_{1n1} \cdot x_{21} \cdot x_{22} \ldots \cdot x_{2n2})^{1/(n1+n2)}$

$$\Rightarrow \log G = \frac{1}{n_1 + n_2} \left(\sum_{i=1}^{n1} \log x_{1i} + \sum_{j=1}^{n2} \log x_{2j} \right) = \frac{1}{n_1 + n_2} (n_1 \log G_1 + n_2 \log G_2)$$

Hence Prove.

1. (a) Define statistics. Explain the scope of statistics.

Ans: Defination of Statistics: Statistics has been defined differently by different authors from time to time. The reasons for a variety of definitions are primarily two. *First,* in modem times the fIeld of utility of Statistics has widened considerably. In ancient times Statistics was confined only to the affairs of State but now it embraces almost every sphere of human activity. Hence a number of old definitions which'were confined to a very narrow field of eQquiry, were replaced by new definitions which are

much more cOl1)prehensive and exhaustive. *Secondly*, Statistics has been defined in two ways. Some writers define it as' *statistical data', i.e.,* numerical statement of facts,while others define itas *'statistical methods', i.e.,* complete body of the principles and techniques used inrcollecting and analysing suc.h data. Some of the i~portant definitions are given below.

Statistics as 'Statistical Data'

Webster defines Statistics ali "classified facts represt;nting the conditions of the people in a State ... especially those facts which can be stated in numbers or in any other tabular or classified arrangement." This definition, since it confines Statistics only to the data pertaining to State; is inadequate as the domain of .Statistics is much wider.

Bowley defines Statistics as " numerical statements of facts in any department of enquiry placed in· relation to each other." A more exhaustive definition is given by Prof. Horace Secrist as follows:

" By Statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre"lletermined purpose and placed in relation tv each other."

Statistics as Statistical,Methods

Bowley himselr d.efines Statistics in. the rollQwing three different ways:

(i) Statistics may be called the science of counting.

(ii) Statistics may rightly be called the science of.averages.

(iii) Statistics is the science of the measurement of social organism, regarded all a whole in all its manifestations.

But none of the above definitions is adequate. The *first* because1tatisticsJis not merely confined to the collection of data as other aspects like presentation, analysis and interpretation, etc., are also covered by it. The *second,* because averages are onl y a part ofthe statistical tools used in the analysis of the data, others' being dispersion, skewness, kurtosi'S, correlation, regression, etc. J'he *third,* because it restricts the application of Statistics to sociology alone while in modem days Statistics is used in almost all sciences - social as well as physical. According to Boddington, " *Statistits is the.. science of estimates and probabilities."* This also is an inadequate definition smce probabilities'and estimates] constitute only a part of the statistical methods.

scope of statistics

In modern times, statistics is viewed not as a mere device for collecting numerical data but as a means of developing sound techniques for their handling and analysis and drawing valid inferences from them. As such it is not confined to the affairs of the state but is intruding constantly into various diversified spheres of life – social as well as physical – such as biology, psychology, education, economics business management etc.

We now discuss briefly the importance of statistics in some different sectors and disciplines.

Statistics and Planning: Statistics is indispensable to planning. In the modern age which is termed as "the age of planning", almost all organizations in the government or managements of business are resorting to planning for efficient working and for formulating policy decision. To achieve this end, the statistical data relating to production, consumption, prices, investment, income, expenditure, etc and various advance statistical techniques for handling and analyzing such complex data are of paramount importance.

Statistics and Mathematics: Statistics is intimately related to and essentially dependent upon mathematics. The modern theory of statistics has its foundations on the theory of probability which in turn is a particular branch of more advanced mathematical theory of measure and Integration. The main stalwarts in the theory of modern statistics namely Laplace, Bernoulli, Pascal, De-Moivre. Even increasing role of mathematics in statistics has resulted in the development of a new branch of Statistics called "*Mathematical Statistics*".

Statistics and Economics: Statistical data and techniques of statistical analysis have proved immensely useful in solving a variety of economic problem, such as wages, process, consumption, production, distribution of income and wealth etc. Statistical tools like's Index numbers, Time series Analysis, Demand Analysis and Forecasting Technique are extensively used for efficient planning and economics development of a country.

Statistics and Business: Statistics is an indispensable tool of production control also. Business executives are relying more and more on statistical techniques for studying the needs and the desires of the consumers and for many other purposes. The success of businessman more or less depends upon the accuracy and

precision of his statistical forecasting. Wrong expectation, which may be the result of faulty and inaccurate analysis of various causes affecting a particular phenomenon, might lead to its disaster.

Statistical techniques have also been used widely by business organization in:

(vii) Carrying out Time and Motion studies.

(viii) Investment (based on the statistical analysis of consumer preference studies – demand analysis).

(ix) Personal Administration (for the study of statistical data relating to wages, cost of living, incentive plans, effect of labor dispute/ unrest on the production, performance standards etc).

(x) Credit Policy

(xi) Inventory Control

(xii) Dale control

Statistical and Biology, Astronomy and Medical Science: The association between statistical methods and biological theories was first studies by Francis Galton in his work in Regression.

(b) What is frequency distribution? Explain its different graphic representation methods with one example.

Ans: FREQUENCY DISTRIBUTION

When observations, discrete or continuous, are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information of interest. Let us consider the marks in statistics obtained by 250 candidates selected at random from among those appearing in a certain examination.

Table (a): Marks in Statistics of 250 Candidates'

32	47	41	51	41	30	39	18	48	53
54	32	31	46	15	37	32	56	42	48
38	26	50	40	38	42	35	22	62	51
44	21	45	31	37	41	44	18	37	47
68	41	30	52	52	60	42	38	38	34
41	53	48	21	28	49	42	36	41	29
30	33	37	35	29	37	38	40	32	49
43	32	24	38	38	22	41	50	17	46
46	50	26	15	23	42	25	52	38	46

41	38	40	37	40	48	45	30	28	31
40	33	42	36	51	42	56	44	35	38
31	51	45	41	50	53	50	32	45	48
40	43	40	34	34	44	38	58	49	28
40	45	19	24	34	47	37	33	37	36
36	32	61	30	44	43	50	31	38	45
46	40	32	32	44	54	35	39	31	48
48	50	43	55	43	39	41	48	53	34
32	31	42	34	34	32	33	24	43	39
40	50	27	47	34	44	34	33	47	42
17	42	57	35	38	17	33	46	36	23
48	50	31	58	33	44	26	29	31	37
47	55	57	37	41	54	42	45	47	43
37	52	47	46	44	50	44	38	42	19
52	45	23	41	47	33	42	24	48	39
48	44	60	38	38	44	38	43	40	48

This representation of the data does not furnish any useful information and is rather confusing to mind. A better way may be to express the figures in an ascending or descending order of magnitude, commonly termed as array. But this does not reduce the bulk of the data. A much better representation is given in Table (b).

Table (b)

Marks	No. of Students -Tally Marks	Total Frequency	Marks	No. of Students -Tally Marks	Total Frequ
15	||	=2	40	𝍸 𝍸 |	=11
17	|||		41		
18	||	=3	42	𝍸 𝍸	=10
19	||		43		
21	||	=2	44	𝍸 𝍸 |||	=13
22			45		
23	||	=2	46	𝍸 |||	=8
24			47		
25	|||	=2	48	𝍸 𝍸 ||	=12
26			49		
27	||||	=2	50	𝍸 ||	=7
28			51		
29	|	=3	52	𝍸 ||	=7
30			53		
31	|||	=4	54	𝍸 |||	=8
32			55		
33	|	=1	56	𝍸 𝍸 ||	=12
34			57		
35	|||	=4	58	||||	=4
36			60		
37	||	=1	61	𝍸 𝍸	=10
38			62		
	𝍸 |	=4		||||	=4

39	𝍸 𝍸	=2	68	𝍸	=5
	𝍸 𝍸	=5		||||	=4
	𝍸 |||	=10		||||	=4
	𝍸 𝍸 |	=10		||	=2
	𝍸	=8		||	=2
	𝍸 𝍸	=11		||	=2
	𝍸 𝍸 ||	=5		||	=2
	𝍸 𝍸 𝍸 ||	=10		|||	=3
	𝍸 |	=12		|	=1
		=17		|	=1
		=6		|	=1

A bar (|) called tally marks is put against the number when it occurs. Having occurred four times, the fifth occurrence is represented by putting a cross tally (|) on the first four tallies.

The representation of the data as above is known as frequency distribution. Marks are called the variable(x) and the number of students against the marks is known the frequency (f) of the variable. The frequency is derived from how frequently a variable occurs. For example, in the above case the frequency of 31 is 10 as there are ten students getting 31 marks.

If the identity of the individuals about whom particular information is taken is not relevant, nor the order in which the observation arise, then the first real step of condensation is to divide the observed range of variable into a suitable number of class-intervals and to record the number of observations in each class. For example, in the above case, the data may be expressed as shown in Table (c)

Table (C)

Marks (x)	No. of students (f)
15 – 19	9
20 – 24	11
25 – 29	10
30 – 34	44
35 – 39	45
40 – 44	54
45 – 49	37
50 – 54	26
55 – 59	8
60 – 64	5

Such a table showing the distribution of the frequencies in the different classes is called a frequency table and the manner in which the class frequencies are distributed over the class intervals is called the grouped frequency distribution of the variable.

The following points may be kept in mind for classification:

1. The classes should be clearly defined and should not lead to any ambiguity.
2. The classes should be exhaustive, i.e., each of the given values should be included in one of the classes.
3. The classes should be mutually exclusive and non-overlapping.

4. The classes should be of equal width. the principle, however, cannot be rigidly followed. If the classes are varying width, the different class frequencies will not be comparable. Comparable figures can be obtained by dividing the value of the frequencies by the corresponding widths of the class intervals. The rations thus obtained are called frequency densities.
5. Indeterminate classes, e.g., the open-end classes like less than 'a' or greater than 'b' should be avoided as far as possible since they create difficulty in analysis and interpretation.
6. The number of classes should neither be too large nor too small. It should preferably lie between 5 and 15.

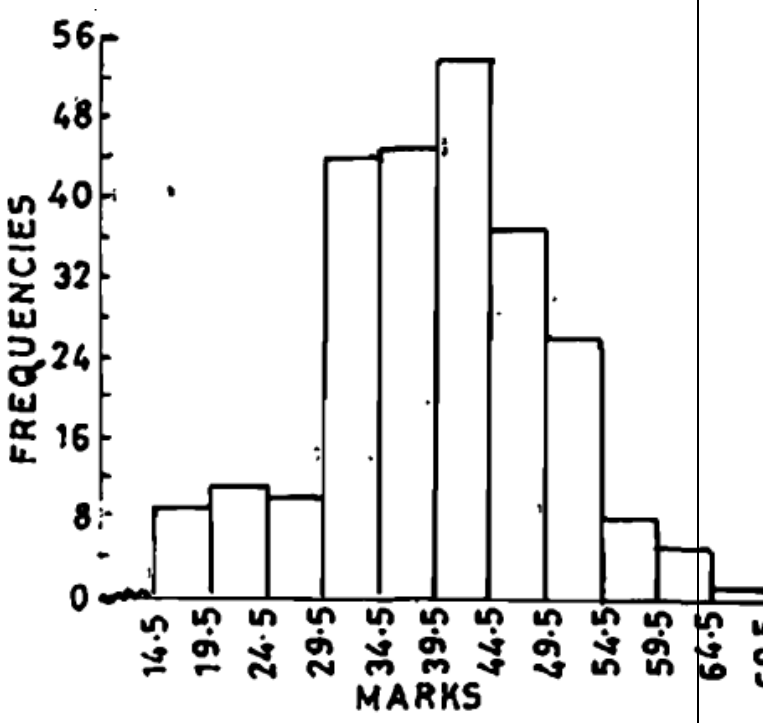
Different Graphics representation Methods

Histogram:

In drawing the histogram of a given continuous frequency distribution we first mark off along the x-axis all the class interval on a suitable scale. On each class interval erect rectangles with heights proportional to the frequency of the corresponding class interval so that the area of the rectangle is proportional to the frequency of the class. If however, the classes are of unequal width then die height of the rectangle will be proportional to the ratio of the frequencies to the width of the classes. The diagram of continuous rectangles so obtained is called histogram.

Since the grouped frequency distribution is not continuous, we first convert it into a continuous distribution as follows:

Marks	No. of Students
14·5–19·5	9
19·5–24·5	11
24·5–29·5	10
29·5–34·5	44
34·5–39·5	45
39·5–44·5	54
44·5–49·5	37
49·5–54·5	26
54·5–59·5	8
59·5–64·5	5
64·5–69·5	1



HISTOGRAM FOR FREQ. DISTRIBUTION

Frequency Polygon

For an ungrouped distribution, the frequency polygon is obtain by plotting points with abscissa as the variate values and the ordinate as the corresponding frequencies and joining the plotted points by means of straight lines. For a grouped frequency distribution, the abscissa of points is mid-values of the class intervals. For equal class intervals the frequency polygon can be obtained by joining the middle Points of the upper sides of the adjacent rectangles of the histogram by means of straight lines. If the class intervals are of small width the polygon can be approximated by a smooth curve. The frequency curve can be obtained by drawing a smooth freehand curve through the vertices of the frequency polygon.

(c) Explain the primary and secondary methods of data collection.

Ans:

Collection of data

Primary and Secondary Data:

Primary data are those which are collected for the first time and are always given in the form of raw materials and originals in character. These types of data need the application of statistics methods for the purpose of analysis and interpretation. While secondary data are

those which have already been collected by someone and have gone thought the statistical machines. They are usually refined of the raw materials .when statistical methods are applied on primary their shape and become secondary data.

Methods of Collection of Primary Data:

The primary data are collected by the following methods.

1. Direct personal investigation.
2. indirect personal investigation
3. Investigation thought questionnaire.
4. investigation through questionnaire in the charge if enumerator
5. Investigation through local's reports.

1. **Direct Personal Investigation:** According to this methods the investigator has to collect his information himself personally form the source concerned. It means the investigator should be are the spot where the enquiry concerned. It means the investigator should be at the spot where the enquiry is being conducted, it is also expected that the investigator should be very polite and courteous. Further he should acquaint himself with the surrounding situation and must know their local customs and tradition.

Advantages:

1. the information collected by this methods is reliable and accurate
2. it is a good methods for intensive investigation
3. This method gives a satisfactory result provided the scope of inquiry is narrow.

Disadvantages:

1. this methods is not suitable for extensive inquiry
2. its required a lot of expenses and time
3. the bias on the part of investigator can damage the whole inquiry
4. sometimes the informant may be reluctant to answered the question

2. **Indirect Personal Investigation:** This method is used when the informants are reluctant to give the definite information. e.g., if a government servant is asked to give the information regarding his income. He will not be willing to give the information for the additional income which he earned by doing part time worked. In such cases what is done? The investigators puts the informant some suitable indirect question which provides him some suitable information. Thus the only difference between the first and the second methods is that in the first methods he investigator puts directs question and collect the information while in the second methods no direct question is put to the informant but only indirect questions are asked. Even then, if it is not possible for the investigator to collect the information by the above methods then the information is collected through indirect sources, i.e. from the persons who have full knowledge of the problem under study. The persons from whom the desired information is collected are known as witnesses. Usually a list of question is prepared which is put before the collected by this method largely depends upon the persons who are selected to give information. Hence it is necessary to take the following precautions for the selection of the informant.

3. **Investigation through questionnaire:** According to this method a standard list of questions relating to the particular investigation is prepared. This list of questions is called a questionnaire. The data are collected "By sending the questionnaire to the informants and requesting them to return the questionnaire after answering the questions. " This method is an important one and is usually used by research workers, non-official bodies and private individuals.

Choice of Questionnaire: The success of the investigation largely depends upon the proper choice the questions to be put to the informants. While preparing a questionnaire the following points should be kept in mind.

i. **Short and clear: -** The questions should be short and clear so as to be easily intelligible to every man. There should be no ambiguity in the questions. If some technical terms are used in the questionnaire, their definitions should be given.

ii. **Few in number and easy:** The questions should be few in number. A large number of the questions would harm the informants because they take much time to answer, with the result they would not pay much attention to ever question and would try to save their akin by giving vague answers. Moreover the questions should be easy to answer.

a. Definiteness: The questions should be such the answers of which are definite and exact. Preferably the questions should be such the replies of which are in the form of "Yes" or "No" Such questions should not be framed the replies of which are vague in nature because such replies are of no use to a statistician.

b. Corroborating in nature: The questions should be such that their replies check the value replies and truth can be easily verified from them.

c. Non-confidential information: The questions framed should not be such which call the confidential information of the informants. This will injure the feelings with the result that they would not give proper answer.

d. Logical sequence: The questions framed should be put in some logical order; their replies should also be put in the same order because this would facilities the work.

4. **Investigation through questionnaire in charge of enumerators:** According into this method enumerators are appointed who go to the informants with the questionnaire and help them in recording the answer. Here the enumerators explain the background, aim and object of the problem under investigation and emphasize the necessity of giving correct answer. They also help the informants in understanding some t4echnical terms of question the concept of which is not clear to the informants. Thus the questionnaire is filled by the informants in the presence and help of the enumerators.

5. **Investigation through local reports:** According to this method the collection of data is neither through the questionnaire nor through the enumerators but through local correspondents. This method of collecting the data is not reliable and it should be used only at those places where the purpose the investigation is served by rough estimates.

COLLECTION OF SECONDARY DATA

The secondary data are those which have already been collected by someone other than the investigator himself, and as such the problems associated with the original collection of data do not arise here. The secondary data can be collected directly either form published or unpublished sources. The following are the sources of published at from which secondary data can be collected.

1. Official publications, i.e. the publication of the central statistical office, Karachi , Ministry of Finance , Ministry of Food, Agriculture, Lahore, Industry, etc… the provincial statistical Bureau, etc.

2. Semi-Official publications , etc., the publication issued by the state Bank of Pakistan Railway Board , Board of Economic Enquiry , District councils, Municipalities, Central Cotton Committee, etc

3. Publication of trade-association, chambers of commerce, co-operative societies, and unions.

4. Research publication, submitted by research workers, economists, University bureaus, and other institutions.
5. Technical or trade journals.

Sources of Unpublished Data: The secondary data are also available from the unpublished data. Type of material can be obtained from the chamber of commerce, trade associations, labor bureaus and research workers.

Scrutiny of Secondary Data: In the words of Bowley, " It is never safe to take published statistics at their face value without knowing their meaning and limitations and it is always necessary to criticize arguments that can be based on them , " Thus the data collected by some other person should not be fully depended as they might have pitfalls. Thus it becomes necessary to find out the inconsistencies probable errors and omissions in the data. This necessities the scrutiny of secondary data because it is just possible that the data might be inaccurate, inadequate or even unsuitable for the purposes of investigation. Hence the secondary data should possess the following qualities:

1. Reliability
2. Suitability
3. Adequacy

1. **RELIABILITY:** In order to test the reliability of the data following points should be considered:

i. Who collected the data?

ii. The source of collection of the data

i. Is the reliability of the compiler dependable?

ii. Is the source of the collection of the data dependable?

iii. What was the scope and object of the investigation?

iv. Were the data collected by the use of proper methods?

v. Were the statistical units defined in which the compiler collected the data?

vi. What was the period of the collection of data?

vii. What was the type of inquiry? Was it census or sample?

viii. What was the degree of accuracy desired and achieved?

ix. Were the data in comparable form?

2. **SUITABILITY:** If the data are reliable it does not mean that they are suitable for every investigation. Data which are found suitable for one inquiry might not be suitable for another one. These necessities that the suitability of the data for the inquiry under investigation is very essential.

(d) Explain the measures of Central tendency with an example.
Ans: Measures of Central Tendency

According to Professor Bowley, averages are "statistical constants which enable us to comprehend in a· single effort the significance of the whole." They give us an idea about the concentration of the values in the central part of the distribution. Plainly speaking, an average of a statistical series is the value of the variable which is representative of the entire distribution. The following are the five measures of Central tendency that are in common use:

(i) **Arithmetic Mean or simply mean**

(ii) **Median**

(iii) **Mode**

(iv) **Geometric Mean and**

(v) **Harmonic Mean**

The following are the characteristics to be satisfied by-an ideal measure of central tendency:

(i) It should be rigidly defined.

(ii) It should be readily comprehensible and easy to calculate.

(iii) It should be based on all the observations.

(iv) It should be suitable for further mathematical treatment. By this we mean that if we are given the averages and sizes of a number of series. We should be able to calculate the average of the composite series obtained on combining the given series.

(v) It should be affected as little as possible by fluctuations of sampling.

(i) **Arithmetic Mean**

Arithmetic mean of a set of observations is their sum divided by the number of observations, e.g. the arithmetic mean $\bar{x}$
of n observation x1,x2,……,xn is given by

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n \right) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

In case of frequency distribution $x_i | f_i$, I = 1,2,….n, where f_i is the frequency of the variable x_i,

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i, \quad \left[\sum_{i=1}^{n} f_i = N \right]$$

In case of grouped of continuous frequency distribution, x is taken as the mid value of the corresponding calss.

Example: Find the arithmetic mean of the following distribution:

X:	1	2	3	4	5	6	7
F:	5	9	12	17	14	10	6

Solution:

X		
1		
2		
3		
4		
5		
6		
7		
	73	**299**

$$\therefore \bar{x} = \frac{1}{N}\sum fx = \frac{299}{73} = 4.09$$

(ii) Median

Median of a distribution is the value of the variable which divides it into two equal pans. It is the v'a1ue which exceeds and is exceeded by the same number of observations, *i.e.*, it ,is the value such that the number of observations above it is equal to the number of observations below it The median is thus a *positional* average.

In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude. In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms.

For example, the median of the values of 25,20,15,35,18, i.e. 15,18,20,25,35,is 20 and the median of 8,20,50,25,15,30,i.e. 8,15,20,25,30,50,is ½(20+25) = 22.5

(iii) Mode: Mode is the value which occurs most frequently in a set of observations and around which the other items of the set cluster densely. In other words, mode is the value of the variable which is predominant in the series. This in the case of discrete frequency distribution mode is the value of x corresponding to maximum frequency. For example in the following frequency distribution:

x	:	1	2	3	4	5	6	7	8
f	:	4	9	16	25	22	15	7	3

the value of x corresponding to the maximum frequency, viz., 25 is 4. Hence mode is 4.

(iv) Geometric Mean: Geometric mean of a set of n observations is the nth root of their product. Thus the geometric mean G of n observations xi, i= 1,2,….n is

$$G = (x1.x2 \ldots . x_n)^{1/n}$$

The computation is facilitated by the use of logarithm. Taking logarithm of both sides, we get

$$\therefore \log G = \frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n) = \frac{1}{n} \sum_{i=1}^{n} \log$$

$$\therefore \quad G = \text{Antilog}\left[\frac{1}{n} \sum_{i=1}^{n} \log x_i\right]$$

In case of frequency distribution xi|fi, (i=1,2,….n) geometric mean, G is given by

$G = [x1^{f1}.x2^{f2} \ldots . xn^{fn}]^{\frac{1}{N}}$, where $N = \sum_{i=1}^{n} fi$

Taking logarithm of both sides, we get

$$\log G = \frac{1}{N}(f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n)$$

$$= \frac{1}{N} \sum_{i=1}^{n} f_i \log x_i$$

Thus we see that logarithm of G is the arithmetic mean of the logarithms of the given values. From equation we get

$$G = \text{Antilog}\left(\frac{1}{N} \sum_{i=1}^{n} f_i \log x_i\right)$$

(v) Harmonic Mean: Harmonics mean of a number of observations is the reciprocal of the arithmetic mean of the reciprocal of the given values. Thus, harmonic mean H, of n observations xi, i= 1,2,….n is

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(1/x_i)}$$

In case of frequency distribution xi|fi, (i=1,2,….,n),

$$H = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}(fi/x_i)}$$

Q1. (a) Discuss the characteristics, rules, limitations and types of Diagrammatic representation of data in detail.

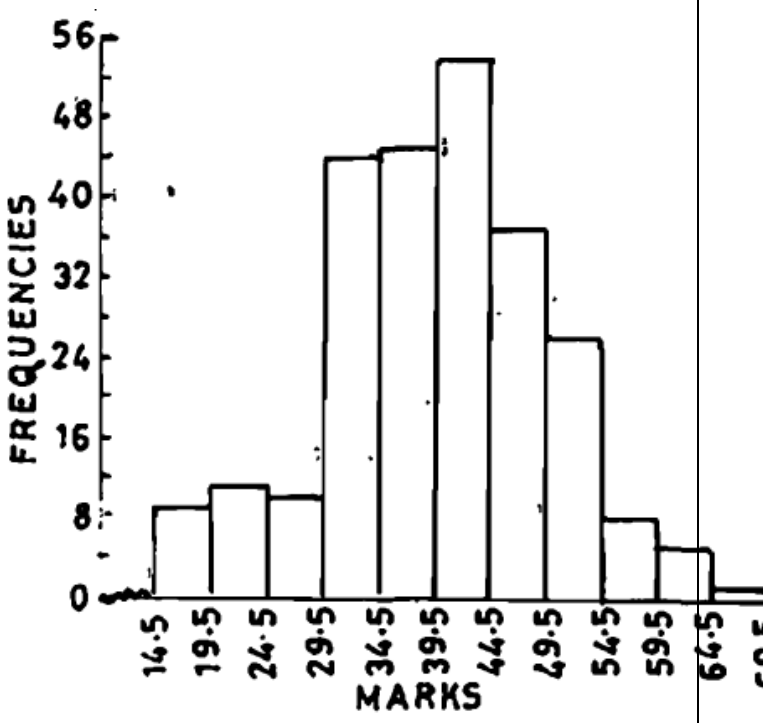
Ans:- The characteristics, rules, limitations and types of Diagrammatic representation of data in detail :-

Histogram:

In drawing the histogram of a given continuous frequency distribution we first mark off along the x-axis all the class interval on a suitable scale. On each class interval erect rectangles with heights proportional to the frequency of the corresponding class interval so that the area of the rectangle is proportional to the frequency of the class. If however, the classes are of unequal width then die height of the rectangle will be proportional to the ratio of the frequencies to the width of the classes. The diagram of continuous rectangles so obtained is called histogram.

Since the grouped frequency distribution is not continuous, we first convert it into a continuous distribution as follows:

Marks	No. of Students
14·5–19·5	9
19·5–24·5	11
24·5–29·5	10
29·5–34·5	44
34·5–39·5	45
39·5–44·5	54
44·5–49·5	37
49·5–54·5	26
54·5–59·5	8
59·5–64·5	5
64·5–69·5	1



HISTOGRAM FOR FREQ. DISTRIBUTION

Frequency Polygon

For an ungrouped distribution, the frequency polygon is obtain by plotting points with abscissa as the variate values and the ordinate as the corresponding frequencies and joining the plotted points by means of straight lines. For a grouped frequency distribution, the abscissa of points is mid-values of the class intervals. For equal class intervals the frequency polygon can be obtained by joining the middle Points of the upper sides of the adjacent rectangles of the histogram by means of straight lines. If the class intervals are of small width the polygon can be approximated by a smooth curve. The frequency curve can be obtained by drawing a smooth freehand curve through the vertices of the frequency polygon.

(B) Derive median formula for continuous frequency distribution and find the median for the following frequency distribution :

Xi	2000 3000	3000 4000	4000 5000	5000 6000	6000 7000
f_i	3	5	20	10	5

Ans:- Derivation of the Median Formula

Let us consider the following continuous frequency distribution, ($x_1 < x_2 < \ldots\ldots < x_{n+1}$):

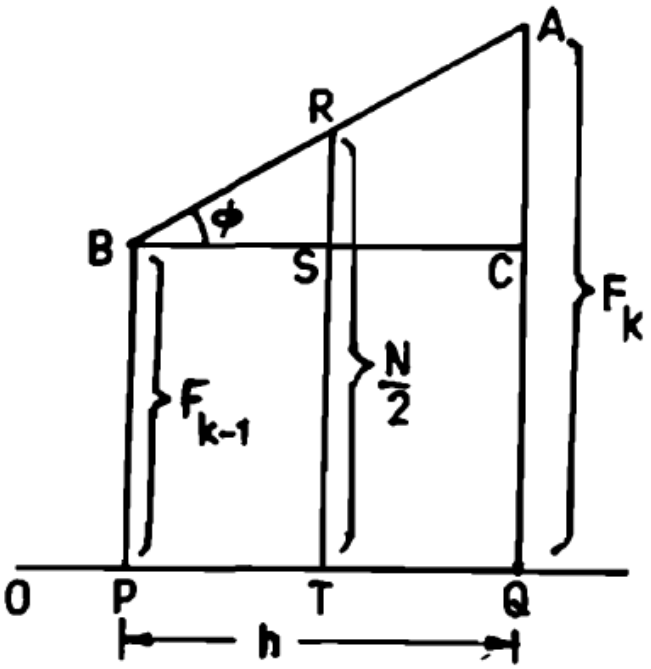
Class interval:	$x_1 - x_2$	$x_2 - x_3$	.....	$x_k - x_{k+1}$	.....	$x_n - x_{n+1}$
Frequency :	f_1	f_2	.....	f_k	.....	f_n

The cumulative frequency distribution is given by:

Class interval:	$x_1 - x_2$	$x_2 - x_3$	.....	$x_k - x_{k+1}$	.....	$x_n - x_{n+1}$
Frequency:	F_1	F_2	.....	F_k	.....	F_n

where $F_i = f_1+f_2+ \ldots\ldots\ldots+f_i$. The class $x_k - x_{k+1}$ is the median class if and only if $F_{k-1} < 1/2N < F_k$

Now, if we assume that the variate values are uniform distributed over the median class which implies that the ogive is a straight line in the median class, then we ger form the fig.



$$\tan \emptyset = \frac{RS}{BS} = \frac{AC}{BC}$$

i.e. $$\frac{RT-TS}{BS} = \frac{AQ-CQ}{BC}$$

or $$\frac{RT-BP}{BS} = \frac{AQ-BP}{PQ}$$

or $$\frac{\frac{1}{2}N - F_{k-1}}{BS} = \frac{F_k - F_{k-1}}{PQ} = \frac{F_k}{h}$$

where f_k is the frequency and h the magnitude of the median class.

$$BS = \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)$$

Hence

Median – OT = OP + PT = OP + BS = l + $\frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)$, which is the required formula.

Xi	2000 – 3000	3000 – 4000	4000 –5000	5000– 6000	6000– 7000
f_i	3	5	20	10	5

Xi	**Fi**	**C.F.**
2000 – 3000	3	3
3000 – 4000	5	8
4000 –5000	20	28
5000– 6000	10	38
6000– 7000	5	43

Here N/2 = 43/2 = 21.5. Cumulative frequency just greater than 21.5 is 28 and corresponding class 4000 – 5000. Thus median class is 4000 – 5000. Hence using formula we get

$$\text{Median} = l + \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)$$

Median = 4000 + 1000 /20 (21.5 – 8)
Median = 4000 + 675
Median = 4675

(c) What do you mean by measures of Central Tendency? Write down the properties of good
measures of Central Tendency. Define Harmonic mean and Geometric mean.
Ans:-
Measures of Central Tendency

According to Professor Bowley, averages are "statistical constants which enable us to comprehend in a· single effort the significance of the whole." They give us an idea about the concentration of the values in the central part of the distribution. Plainly speaking, an average of a statistical series is the value of the variable which is representative of the entire distribution. The following are the five measures of Central tendency that are in common use:

(i) **Arithmetic Mean or simply mean**

(ii) **Median**

(iii) **Mode**

(iv) **Geometric Mean and**

(v) **Harmonic Mean**

The following are the characteristics to be satisfied by-an ideal measure of central tendency:

(i) It should be rigidly defined.

(ii) It should be readily comprehensible and easy to calculate.

(iii) It should be based on all the observations.

(iv) It should be suitable for further mathematical treatment. By this we mean that if we are given the averages and sizes of a number of series. We should be able to calculate the average of the composite series obtained on combining the given series.

(vi) It should be affected as little as possible by fluctuations of sampling.

(vii) It should not affected by extreme values.

Geometric Mean: Geometric mean of a set of n observations is the nth root of their product. Thus the geometric mean G of n observations xi, i= 1,2,….n is

$$G = (x1.x2 \ldots . x_n)^{1/n}$$

The computation is facilitated by the use of logarithm. Taking logarithm of both sides, we get

$$\therefore \log G = \frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n) = \frac{1}{n}\sum_{i=1}^{n} \log$$

$$\therefore \quad G = \text{Antilog}\left[\frac{1}{n}\sum_{i=1}^{n} \log x_i\right]$$

In case of frequency distribution xi|fi, (i=1,2,….n) geometric mean, G is given by

$G = [x1^{f1}.x2^{f2}\ldots.xn^{fn}]^{\frac{1}{N}}$, where $N = \sum_{i=1}^{n} fi$

Taking logarithm of both sides, we get

$$\log G = \frac{1}{N}(f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n)$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \log x_i$$

Thus we see that logarithm of G is the arithmetic mean of the logarithms of the given values. From equation we get

$$G = \text{Antilog}\left(\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right)$$

Harmonic Mean: Harmonics mean of a number of observations is the reciprocal of the arithmetic mean of the reciprocal of the given values. Thus, harmonic mean H, of n observations xi, i= 1,2,….n is

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(1/x_i)}$$

In case of frequency distribution xi|fi, (i=1,2,….,n),

$$H = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}(fi/x_i)}$$

(d) Find the mode for the following distribution using grouping method:

X_i	1	2	3	4	5	6	7	8	9	10	11	12
F	3	8	15	23	35	40	32	28	20	45	14	6

Sol:- Here we see that the distribution is not regular since the frequencies are increasing steadily up to 40 and then decrease but the frequency 45 after 20 does not seem to be consistent with the distribution. Here we cannot say that since maximum frequency is 45, mode is 10. Here we shall locate mode by the method of grouping explained below:

The frequencies in column (i) are the original frequencies. Column (ii) is obtained by combining the frequencies two by two. If we leave the first frequency and combine the remaining frequencies two by two get column (iii). Combining the frequencies two by two after leaving the first two frequencies results in a repetition of column (ii). Hence, we proceed to combine the frequencies three by three, thus getting column (iv). The combination of frequencies three by three after leaving the first frequency result in column (v) and after leaving the first two frequencies result in column (vi).

Size (x)	Frequency (i)	(ii)	(iii)	(iv)	(v)	(vi)
1	3					
2	8	11				
3	15		23			
4	23	38		26		
5	35		58			
6	40	75			46	
7	32		72			73
8	28	60		98		
9	20		48			
10	**45**	65			107	
11	14		59	80		100
12	6	20				
				65		
					93	
						79

The maximum frequency in each column is given in black type. To find the mode we form the following table:

ANALYSIS TABLE

Column Number **(1)**	**Maximum Frequency** **(2)**	**Value or combination of values of x giving max frequency in (2) (3)**
(i)	45	10
(ii)	75	5, **6**
(iii)	72	**6**, 7
(iv)	98	4,5,**6**
(v)	107	5,**6**,7
(vi)	100	**6**,7,8

On examining the values in column (3) above, we find that the value 6 is repeated the maximum number of times and hence the value of mode is **6**.

1. **(A) Discuss different methods of collecting Primary data and Secondary data.**

Ans:-

Collection of data

Primary and Secondary Data:

Primary data are those which are collected for the first time and are always given in the form of raw materials and originals in character. These types of data need the application of statistics methods for the purpose of analysis and interpretation. While secondary data are those which have already been collected by someone and have gone thought the statistical machines. They are usually refined of the raw materials .when statistical methods are applied on primary their shape and become secondary data.

Methods of Collection of Primary Data:

The primary data are collected by the following methods.

1. Direct personal investigation.
2. indirect personal investigation
3. Investigation thought questionnaire.
4. investigation through questionnaire in the charge if enumerator
5. Investigation through local's reports.

1. **Direct Personal Investigation:** According to this methods the investigator has to collect his information himself personally form the source concerned. It means the investigator should be are the spot where the enquiry concerned. It means the investigator should be at the spot where the enquiry is being conducted, it is also expected that the investigator should be very polite and courteous. Further he should acquaint himself with the surrounding situation and must know their local customs and tradition.

Advantages:

1. the information collected by this methods is reliable and accurate
2. it is a good methods for intensive investigation
3. This method gives a satisfactory result provided the scope of inquiry is narrow.

Disadvantages:

1. this methods is not suitable for extensive inquiry
2. its required a lot of expenses and time
3. the bias on the part of investigator can damage the whole inquiry
4. sometimes the informant may be reluctant to answered the question

2. **Indirect Personal Investigation:** This method is used when the informants are reluctant to give the definite information. e.g., if a government servant is asked to give the information regarding his income. He will not be willing to give the information for the additional income which he earned by doing part time worked. In such cases what is done? The investigators puts the informant some suitable indirect question which provides him some suitable information. Thus the only difference between the first and the second methods is that in the first methods he investigator puts directs question and collect the information while in the second methods no direct question is put to the

informant but only indirect questions are asked. Even then, if it is not possible for the investigator to collect the information by the above methods then the information is collected through indirect sources, i.e. from the persons who have full knowledge of the problem under study. The persons from whom the desired information is collected are known as witnesses. Usually a list of question is prepared which is put before the collected by this method largely depends upon the persons who are selected to give information. Hence it is necessary to take the following precautions for the selection of the informant.

3. **Investigation through questionnaire:** According to this method a standard list of questions relating to the particular investigation is prepared. This list of questions is called a questionnaire. The data are collected "By sending the questionnaire to the informants and requesting them to return the questionnaire after answering the questions. " This method is an important one and is usually used by research workers, non-official bodies and private individuals.

Choice of Questionnaire: The success of the investigation largely depends upon the proper choice the questions to be put to the informants. While preparing a questionnaire the following points should be kept in mind.

i. **Short and clear: -** The questions should be short and clear so as to be easily intelligible to every man. There should be no ambiguity in the questions. If some technical terms are used in the questionnaire, their definitions should be given.

ii. **Few in number and easy:** The questions should be few in number. A large number of the questions would harm the informants because they take much time to answer, with the result they would not pay much attention to ever question and would try to save their akin by giving vague answers. Moreover the questions should be easy to answer.

a. Definiteness: The questions should be such the answers of which are definite and exact. Preferably the questions should be such the replies of which are in the form of "Yes" or "No" Such questions should not be framed the replies of which are vague in nature because such replies are of no use to a statistician.

b. Corroborating in nature: The questions should be such that their replies check the value replies and truth can be easily verified from them.

c. Non-confidential information: The questions framed should not be such which call the confidential information of the informants. This will injure the feelings with the result that they would not give proper answer.

d. Logical sequence: The questions framed should be put in some logical order; their replies should also be put in the same order because this would facilities the work.

4. **Investigation through questionnaire in charge of enumerators:** According into this method enumerators are appointed who go to the informants with the questionnaire and help them in recording the answer. Here the enumerators explain the background, aim and object of the problem under investigation and emphasize the necessity of giving correct answer. They also help the informants in understanding some t4echnical terms of question the concept of which is not clear to the informants. Thus the questionnaire is filled by the informants in the presence and help of the enumerators.

5. **Investigation through local reports:** According to this method the collection of data is neither through the questionnaire nor through the enumerators but through local correspondents. This method of collecting the data is not reliable and it should be used only at those places where the purpose the investigation is served by rough estimates.

COLLECTION OF SECONDARY DATA

The secondary data are those which have already been collected by someone other than the investigator himself, and as such the problems associated with the original collection of data do not arise here. The secondary data can be collected directly either form published or unpublished sources. The following are the sources of published at from which secondary data can be collected.

1. Official publications, i.e. the publication of the central statistical office, Karachi , Ministry of Finance , Ministry of Food, Agriculture, Lahore, Industry, etc… the provincial statistical Bureau, etc.
2. Semi-Official publications , etc., the publication issued by the state Bank of Pakistan Railway Board , Board of Economic Enquiry , District councils, Municipalities, Central Cotton Committee, etc
3. Publication of trade-association, chambers of commerce, co-operative societies, and unions.
4. Research publication, submitted by research workers, economists, University bureaus, and other institutions.
5. Technical or trade journals.

Sources of Unpublished Data: The secondary data are also available from the unpublished data. Type of material can be obtained from the chamber of commerce, trade associations, labor bureaus and research workers.

Scrutiny of Secondary Data: In the words of Bowley, " It is never safe to take published statistics at their face value without knowing their meaning and limitations and it is always necessary to criticize arguments that can be based on them , " Thus the data collected by some other person should not be fully depended as they might have pitfalls. Thus it becomes necessary to find out the inconsistencies probable errors and omissions in the data. This necessities the scrutiny of secondary data because it is just possible that the data might be inaccurate, inadequate or even unsuitable for the purposes of investigation. Hence the secondary data should possess the following qualities:

1. Reliability
2. Suitability
3. Adequacy

1. **RELIABILITY:** In order to test the reliability of the data following points should be considered:

i. Who collected the data?

ii. ii. The source of collection of the data

i. Is the reliability of the compiler dependable?

ii. Is the source of the collection of the data dependable?

iii. What was the scope and object of the investigation?

iv. Were the data collected by the use of proper methods?

v. Were the statistical units defined in which the compiler collected the data?

vi. What was the period of the collection of data?

vii. What was the type of inquiry? Was it census or sample?

viii. What was the degree of accuracy desired and achieved?

ix. Were the data in comparable form?

2. **SUITABILITY:** If the data are reliable it does not mean that they are suitable for every investigation. Data which are found suitable for one inquiry might not be suitable for another one. These necessities that the suitability of the data for the inquiry under investigation is very essential.

(B) In a factory employing 3000 persons in a day. 5 % work less than 3 hours, 580 works from 3.01 to 4.50 hours, 30 % work from 4.51 to 6.00 hours, 500 work from 6.01 to 7.50 hours, 20 % work from 7.51 to 9.00 hours and rest work 9.01 or more hours. What are the median hours of work?

Ans:- The given information can be expressed in tabular from as follows:

Work Hours	No. of Employees (f)	Less than c.f.	Class boundaries
Less than 3	5/100 * 3000 = 150	150	Below 3.005
3.01 – 4.50	580	730	3.005 – 4.505
4.51 – 6.00	30/100*3000 = 900	1630	4.505 – 6.005
6.01 – 7.50	500	2130	6.005 – 7.505
7.51 – 9.00	20/100*3000= 600	2730	7.505 – 9.005
9.01 and above	3000- 2730 = 270	3000 = N	9.005 and above

Here N = 3000 => ½ N = 1500.

The c.f. just greater than 1500 is 1630. The corresponding class 4.51 – 6.00 is the median class.

Using the median formula, we get

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right) = 4.505 + \frac{1.5}{900}(1500 - 730) = 4.505 + 1.283 = 5.79$$

Hence, the median hours of work are 5.79.

(C) The following is the age distribution of 1000 persons working in a large industrial house :

Age Group	No. of Persons
20—25	30
25—30	160
30—35	210
35—40	180
40—45	145
45—50	105
50—55	70
55—60	60
60—65	40

Due to continuous heavy losses the management decides to bring down the strengths to 30 % of the present number according to the following schemes :
(1) To retrench the first 15 % from the lower group.
(2) To absorb the next 45 % in other branch.
(3) To make 10 % from the highest age group retired permanently.
Calculate the age limits of the persons retained and those to be transferred to other departments.
Also find the average age of those retained.
Ans: Total number of persons in the industrial house is N = 1000.
According to the conditions of the problem:

(i) The number of persons to be retrenched from the lower group
= 15% of N = 15/100 * 1000 = 150
30 of these will be from the first group 20 – 25 and the remaining 150 – 30 = 120 from the next age group 25 – 30.

(ii) The number of persons (from the next groups) to be absorbed in other branches:
= 45 % of N = 45/100 * 1000 = 450.
These will belong to the different age group as detailed below:

Age Group	**No. of Persons**
25—30	160 – 120 = 40
30—35	210
35—40	180
40—45	450 – (40 + 210 + 180) = 20

(iii)The number of persons to retire (from the highest age – groups)
= 10 % of N = 10/100 * 1000 = 100.
These 100 persons are from the highest age group as shown:

Age Group	**No. of Persons**
55 – 60	60
60 – 65	40

Hence incorporating the steps in (i), (ii) and (iii) the frequency distribution of the number of persons retained in the industrial house is as shown in the adjoining table:

Age Group	**No. of Persons**
40 – 45	145 – 20 =125
45 – 50	105
50 – 55	70

Calculation for average age of those retained

Age Group	Mid value (x)	F	$d = \frac{x - 47.5}{5}$	fd
40 – 45	42.5	125	-1	-125
45 – 50	47.5	105	0	0
50 – 55	52.5	70	1	70
Total	N = 300		$\sum fd = -55$	

The average age of those retained is given by:

$$\bar{X} = A + \frac{h\sum fd}{N} = 47.5 + 5 \times \frac{(-55)}{300} = 46.5833 \approx 47$$

Hence, the average age of those retained in the industrial house is 47 years.

(D) Define frequency distribution. Explain different types of frequency distribution. What are different graphs used for representing continuous frequency distribution?

Ans:- : FREQUENCY DISTRIBUTION

When observations, discrete or continuous, are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information of interest. Let us consider the marks in statistics obtained by 250 candidates selected at random from among those appearing in a certain examination.

Table (a): Marks in Statistics of 250 Candidates'

32	47	41	51	41	30	39	18	48	53
54	32	31	46	15	37	32	56	42	48
38	26	50	40	38	42	35	22	62	51
44	21	45	31	37	41	44	18	37	47
68	41	30	52	52	60	42	38	38	34
41	53	48	21	28	49	42	36	41	29
30	33	37	35	29	37	38	40	32	49
43	32	24	38	38	22	41	50	17	46
46	50	26	15	23	42	25	52	38	46
41	38	40	37	40	48	45	30	28	31
40	33	42	36	51	42	56	44	35	38

31	51	45	41	50	53	50	32	45	48
40	43	40	34	34	44	38	58	49	28
40	45	19	24	34	47	37	33	37	36
36	32	61	30	44	43	50	31	38	45
46	40	32	32	44	54	35	39	31	48
48	50	43	55	43	39	41	48	53	34
32	31	42	34	34	32	33	24	43	39
40	50	27	47	34	44	34	33	47	42
17	42	57	35	38	17	33	46	36	23
48	50	31	58	33	44	26	29	31	37
47	55	57	37	41	54	42	45	47	43
37	52	47	46	44	50	44	38	42	19
52	45	23	41	47	33	42	24	48	39
48	44	60	38	38	44	38	43	40	48

This representation of the data does not furnish any useful information and is rather confusing to mind. A better way may be to express the figures in an ascending or descending order of magnitude, commonly termed as array. But this does not reduce the bulk of the data. A much better representation is given in Table (b).

Table (b)

Marks	No. of Students -Tally Marks	Total Frequency	Marks	No. of Students -Tally Marks	Total Frequ
15	||	=2	40	𝍸 𝍸 |	=11
17	|||		41		
18	||	=3	42	𝍸 𝍸	=10
19	||		43		
21	||	=2	44	𝍸 𝍸 |||	=13
22			45		
23	||	=2	46	𝍸 |||	=8
24			47		
25	|||	=2	48	𝍸 𝍸 ||	=12
26			49		
27	||||	=2	50	𝍸 ||	=7
28			51		
29	|	=3	52	𝍸 ||	=7
30			53		
31	|||	=4	54	𝍸 |||	=8
32			55		
33	|	=1	56	𝍸 𝍸 ||	=12
34			57		
35	|||	=4	58	||||	=4
36			60		
37	||	=1	61	𝍸 𝍸	=10
38			62		
39	𝍸 |	=4	68	||||	=4
	𝍸 𝍸	=2		𝍸	=5

𝍸 𝍸	=5	||||	=4
𝍸 |||	=10	||||	=4
𝍸 𝍸 |	=10	||	=2
𝍸	=8	||	=2
𝍸 𝍸	=11	||	=2
𝍸 𝍸 ||	=5	||	=2
𝍸 𝍸 𝍸 ||	=10	|||	=3
𝍸 |	=12	|	=1
	=17	|	=1
	=6	|	=1

A bar (|) called tally marks is put against the number when it occurs. Having occurred four times, the fifth occurrence is represented by putting a cross tally (||) on the first four tallies.

The representation of the data as above is known as frequency distribution. Marks are called the variable(x) and the number of students against the marks is known the frequency (f) of the variable. The frequency is derived from how frequently a variable occurs. For example, in the above case the frequency of 31 is 10 as there are ten students getting 31 marks.

If the identity of the individuals about whom particular information is taken is not relevant, nor the order in which the observation arise, then the first real step of condensation is to divide the observed range of variable into a suitable number of class-intervals and to record the number of observations in each class. For example, in the above case, the data may be expressed as shown in Table (c)

Table (C)

Marks (x)	No. of students (f)
15 – 19	9
20 – 24	11
25 – 29	10
30 – 34	44
35 – 39	45
40 – 44	54
45 – 49	37
50 – 54	26
55 – 59	8
60 – 64	5

Such a table showing the distribution of the frequencies in the different classes is called a frequency table and the manner in which the class frequencies are distributed over the class intervals is called the grouped frequency distribution of the variable.

The following points may be kept in mind for classification:

7. The classes should be clearly defined and should not lead to any ambiguity.
8. The classes should be exhaustive, i.e., each of the given values should be included in one of the classes.
9. The classes should be mutually exclusive and non-overlapping.
10. The classes should be of equal width. the principle, however, cannot be rigidly followed. If the classes are varying width, the different class frequencies will not be

comparable. Comparable figures can be obtained by dividing the value of the frequencies by the corresponding widths of the class intervals. The rations thus obtained are called frequency densities.

11. Indeterminate classes, e.g., the open-end classes like less than ‘a’ or greater than ‘b’ should be avoided as far as possible since they create difficulty in analysis and interpretation.
12. The number of classes should neither be too large nor too small. It should preferably lie between 5 and 15.

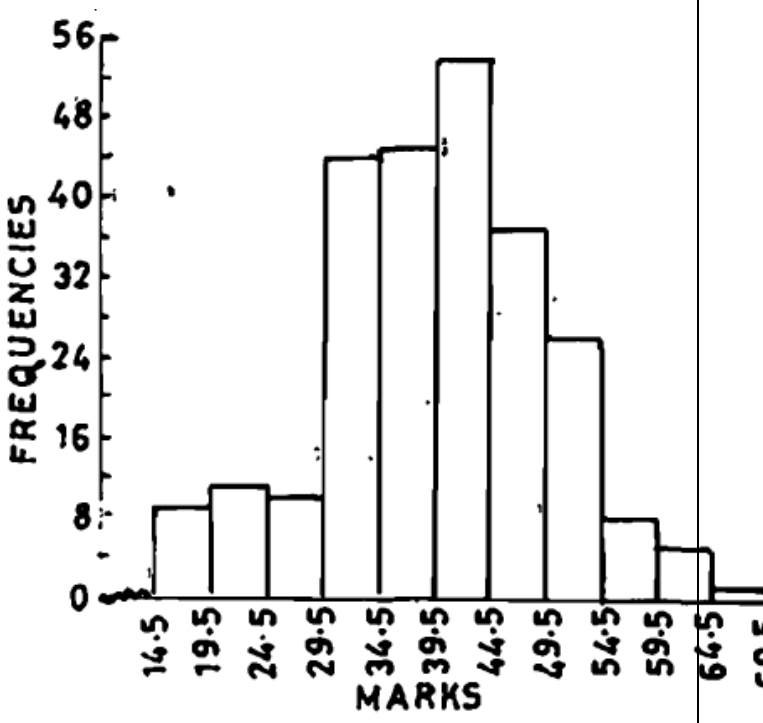
Different Graphics representation Methods

Histogram:

In drawing the histogram of a given continuous frequency distribution we first mark off along the x-axis all the class interval on a suitable scale. On each class interval erect rectangles with heights proportional to the frequency of the corresponding class interval so that the area of the rectangle is proportional to the frequency of the class. If however, the classes are of unequal width then die height of the rectangle will be proportional to the ratio of the frequencies to the width of the classes. The diagram of continuous rectangles so obtained is called histogram.

Since the grouped frequency distribution is not continuous, we first convert it into a continuous distribution as follows:

Marks	*No. of Students*
14·5–19·5	9
19·5–24·5	11
24·5–29·5	10
29·5–34·5	44
34·5–39·5	45
39·5–44·5	54
44·5–49·5	37
49·5–54·5	26
54·5–59·5	8
59·5–64·5	5
64·5–69·5	1



HISTOGRAM FOR FREQ. DISTRIBUTION

Frequency Polygon

For an ungrouped distribution, the frequency polygon is obtain by plotting points with abscissa as the variate values and the ordinate as the corresponding frequencies and joining the plotted points by means of straight lines. For a grouped frequency distribution, the abscissa of points is mid-values of the class intervals. For equal class intervals the frequency polygon can be obtained by joining the middle Points of the upper sides of the adjacent rectangles of the histogram by means of straight lines. If the class intervals are of small width the polygon can be approximated by a smooth curve. The frequency curve can be obtained by drawing a smooth freehand curve through the vertices of the frequency polygon.

UNIT – II

Q(2)

(A) Define correlation coefficient and explain limits of correlation coefficient.

(S-14)

Ans: **Correlation Coefficient:** A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. In positively correlated variables, the value increases or decreases in tandem. In negatively correlated variables, the value of one increases as the value of the other decreases.

Correlation coefficients are expressed as values between +1 and -1. A coefficient of +1 indicates a perfect positive correlation: A change in the value of one variable will predict a change in the same direction in the second variable. A coefficient of -1 indicates a perfect negative correlation: A change in the value of one variable predicts a change in the opposite direction in the second variable. Lesser degrees of correlation are expressed as non-zero decimals. A coefficient of zero indicates there is no discernable relationship between fluctuations of the variables.

Limits for Correlation Coefficient

We have

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{[\frac{1}{n}\sum(x_i - \bar{x})^2 . \frac{1}{n}(y_i - \bar{y})^2]^{\frac{1}{2}}}$$

$$\therefore\ r^2(X,Y) = \frac{(\sum_i a_i b_i)^2}{(\sum_i a_i^2)(\sum_i b_i^2)}, \text{ where } \begin{pmatrix} a_i = x_i - \bar{x} \\ b_i = y_i - \bar{y} \end{pmatrix} \quad \text{.....(*)}$$

We have the Schwartz inequality which states that if a_i, b_i, i = 1,2,.....,n are real quantities then

$(\sum_{i=1}^{n} a_i b_i)^2 \leq (\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)$,

The sign of equality holding if and only if $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \cdots = \frac{a_n}{b_n}$

Using Schwartz inequality, we get from (*):

$$r^2(X,Y) \leq 1 \; i.e. \, |r(X,Y)| \leq 1 \implies -1 \leq r(X,Y) \leq 1$$

Hence correlation coefficient cannot exceed unity numerically. It always lies between -1 and +1.

(B) Prove that standard deviation is independent of change of origin but not scale.

Ans: By definition of standard deviation (S-14)

$$\sigma^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2$$

More precisely we write it as σ_x^2 , $i.e$ variance of x. Thus

$$\sigma_x^2 = f_i (x_i - \bar{x})^2$$

$\bar{x}$ is not a whole number but comes out to be in fraction, the calculation of σ_x^2 by using (1) is very cumbersome and time consuming. In order to overcome this difficulty, we shall develop different forms of the formula (1) which reduce the arithmetic toa great extent and are very useful for computational work. In the following discussion, the summation is extended over i form 1 to n.

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_i f_i (x_i^2 + \bar{x}^2 - 2 x_i \bar{x})$$

$$= \frac{1}{N} \sum_i f_i x_i^2 + \bar{x}^2 \frac{1}{N} \sum_i f_i - 2\bar{x} \cdot \frac{1}{N} \sum_i f_i x_i$$

$$= \frac{1}{N} \sum_i f_i x_i^2 + \bar{x}^2 - 2\bar{x}^2 = \frac{1}{N} \sum_i f_i x_i^2 - \bar{x}^2$$

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i x_i^2 - \left(\frac{1}{N} \sum_i f_i x_i \right)^2$$

If the values of x and f are large, the calculation of fx, fx^2 is quite tedious. In that case we take the deviations from any arbitrary point 'A'. Generally the point in the middle of the distribution is much convenient though the formula is true in general.

Let $d_i = x_i - A$ ..... (*)

Multiplying by fi, summing over i to n, and dividing by N, we get

$\frac{1}{N}\sum_i f_i d_i = \frac{1}{N}\sum_i f_i x_i - A.\frac{1}{N}\sum f_i \quad \Rightarrow \bar{d} = \bar{x} - A$ ... (**)

Subtracting (**) from (*), we get: $d_i - \bar{d} = x_i - \bar{x}$

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(d_i - \bar{d})^2 = \sigma_d^2$$

Hence, variance and consequently standard deviation is independent of change of origin.

If $d_i = \frac{x_i - A}{h}$, then $x_i = A + hd_i$

$\bar{x} = A + h.\frac{1}{N}\sum_i f_i d_i = A + h\bar{d} \quad \Rightarrow x_i - \bar{x} = h(d_i - \bar{d})$

And $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2 = h^2\frac{1}{N}\sum_{i=1}^{n} f_i(d_i - \bar{d})^2 = h^2\sigma_d^2$

Hence, variance is independent of change of origin but not scale.

(c) Find the rank correlation coefficient for the following data: (S – 14)

X	**41**	**42**	**49**	**48**	**50**	**51**	**49**
Y	**10**	**09**	**15**	**17**	**20**	**17**	**18**

Ans: CALCULATION FOR RANK CORRELATION

X	Y	Rank X (x)	Rank Y (y)	D= x-y	D^2
40	10	7	6	1	1
42	09	6	7	-1	1
49	15	3.5	5	-1.5	2.25
48	17	5	3.5	1.5	2.25
50	20	2	1	1	1
51	17	1	3.5	-2.5	6.25
49	18	3.5	2	1.5	2.25
				$\sum d = 0$	$\sum d^2 = 16$

In the X- Series we see that the value 49 occurs 2 times. The common rank given to these values is 3.5 which is the average of 3 and 4, the ranks which these values would have taken if they were different. Similarly in the Y-series, the value 17 occurs twice and its common rank is 3.5 which is the average of 3 and 4. As a result of these common rankings, the formula for 'ρ' has to be corrected. To $\sum d^2$ we add $\frac{m(m^2-1)}{12}$ for each value repeated, where m is the number of times a value occurs. In the X-series the correction is

to be applied once for the values of 49 which occur twice (m=2). The total correction for the X-series is : $\frac{2(4-1)}{12} = \frac{1}{2}$

Similarly, this correction for the Y-series is $\frac{2(4-1)}{12} = \frac{1}{2}$ as the value of 17 occurs twice.

$$\rho = 1 - \frac{6\left(\sum d^2 + \frac{1}{2} + \frac{1}{2}\right)}{n(n^2-1)}$$

$$= 1 - \frac{6(16+1)}{7(49-1)} \quad = 1 - \frac{102}{336} \quad = 1 - 0.30 = 0.7$$

(d) Explain meaning and significance of dispersion. (S – 14)

Ans:- Averages or the measures of central tendency give us an idea of tile concentration of the observation about the central part of the distribution. If we know the average alone we cannot form a complete idea about the distribution as will be clear from the following example.

Consider the series (i) 7. 8, 10, 11, (ii) 3, 6; 9, 12, 15, (iii) I, 5, 9, 13, 17. In all these cases we see that n, the number of observations is 5 and the mean is 9. If we arc given that the mean of 5 observations is 9, we cannot form an idea as to whether it is the average of first series or second series or third series or of any other series (If 5 observations whose sum is 45. Thus we see that the measures of central tendency are inadequate to give us a complete idea of the distribution. They must be supported and supplemented by some other measures,

One such measure is Dispersion.

Literal meaning of dispersion is ‘scatteredness'. We study dispersion to have an idea about the homogeneity or heterogeneity of the distribution. In the above case we say that series (i) is more homogeneous (less dispersed) than the series (ii) Qr (iii) or we say that series (iii) is more heterogeneous (more scattered) than the series (i) or (ii).

2. **(a) What do you mean by regression? Derive an equation for the line of regression of y on x.**

   Ans:-

   Regression: Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.

   In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the values or is used for prediction. is called independent variable. In

regression analysis independent variable is also known as regressor or predictor or explanatory variable while the dependent variable is also known as regressed or explained variable.

Line of Regression: If the variables in a bivariate distribution are related. we will find that the points in the scatter diagram will cluster round some curve called the "curve of regression". If the curve is a straight line, it is called the line of regression and there is said to be linear regression' between the variables, otherwise regression is said to be curvilinear.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of "best fit" and is obtained by the principles of least squares.

Let us suppose that in the bivariate distribution (xi,yi); i=1,2,….,n; Y is dependent variable and X is independent variable. Let the line of regression of Y on X be Y = a+bX.

According to the principle of least square, the normal equations for estimating a and b are

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \qquad \text{……..(1)}$$

And $$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \qquad \text{……..(2)}$$

From (1) on dividing by n, we get

$$\bar{y} = a + b\bar{x}$$

Thus the line of regression Y on X passes through the point $(\bar{x}, \bar{y})$

Now

$$\mu_{11} = COV(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\bar{y}$$

……..(3)

Also $$\sigma_{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x^2} \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma_{x^2} + \bar{x}^2 \qquad \text{……..(3a)}$$

Dividing (2) by n and using (3) and (3a) we get

$$\mu_{11} + \bar{x}\,\bar{y} = a\bar{x} + b(\sigma_x^2 + \bar{x}^2) \qquad \text{…….. (4)}$$

Multiplying by $\bar{x}$ and then subtracting from (4) we get

$$\mu_{11} = b\sigma_x^2 \quad \Rightarrow \quad b = \frac{\mu_{11}}{\sigma_x^2} \qquad \text{……..(5)}$$

Since 'b' is the slope of the line of regression of Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y})$, its equation is

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x})$$

$$\Rightarrow \quad Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

Starting with the equation X= A +BY and proceeding similarly or by simply interchanging the variables X and Y in above equations, the equation of the line of regression of X on Y becomes

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y})$$

$$\Rightarrow \quad X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

(b) Find mean deviation and standard deviation for age distribution of 40 workers

Age (yrs)	20 – 30	30 – 40	40 – 50	50 – 60
No. of workers	4	9	15	12

Ans:

Here we take $d = \frac{x-A}{h} = \frac{x-45}{10}$

Age(yrs)	Mid – value (x)	No. of worker (f)	$d = \frac{x-45}{10}$	fd	Fd^2
20 – 30	25	4	-2	-8	64
30 – 40	35	9	-1	-9	81
40 – 50	45	15	0	0	0
50 – 60	55	12	1	12	144

		N=$\sum f = 40$		$\sum fd = -5$	$\sum fd^2 = 289$

$$\bar{x} = A + h\frac{\sum fd}{N} = 45 + \frac{10 \times (-5)}{40} = 45 - 1.25 = 46.25$$

$$\sigma^2 = h^2\left[\frac{1}{N}\sum fd^2 - \left(\frac{1}{N}\sum fd\right)^2\right]$$
$$= 100\left[\frac{289}{40} - (-0.375)^2\right] = 100 \times 7.085 = 708.5$$
$$\therefore\ \sigma\ (Standard\ deviation) = 26.617$$

(c) Explain different measures of dispersion

Ans: - The following are the measures of dispersions

i) Range,
ii) Quartile deviation or Semi – interquartile range,
iii) Mean Deviation, and
iv) Standard Deviation

i) Range : The range is the difference between two extreme observations, or the distribution. If A and B are the greatest and smallest observations respectively in a distribution, then its range is A-B.
Range is the simples I but a crude measure of dispersion. Since it is based on two extreme observations which themselves are subject to change fluctuations, it is not at all a reliable measure of dispersion.

ii) Quartile Deviation : Quartile deviation or semi – interquartile range Q is given by

$$\mathbf{Q = ½\ (Q_3 - Q_1),}$$

Where Q1 and Q3 are the first and third quartile of the distribution respectively. Quartile deviation is definitely a better measure than the range as it makes use of 50% of the data. But since it ignores the other 50% of the data, it cannot be regarded as a reliable measure.

iii) Mean Deviation : If xi| fi, i=1,2,….,n is the frequency distribution, then mean deviation from the average A, (usually mean, median or mode) is given by

$$\text{Mean deviation} = \frac{1}{N}\sum_i f_i |x_i - A|, \quad \sum f_i = N$$

Where |xi – A| represents the modulus or the obsolute value of the deviation (xi – A), when the –ve signed is ignored.

Since mean deviation is based on all the observations, it is a better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviations (xi – A) create artificiality and renders it useless for further mathematical treatment.

iv) **Standard Deviation :** Standard deviation, usually denoted by the Greek letter small sigma (σ), is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. For the frequency distribution xi|fi, i= 1,2,…,n

$$\sigma = \sqrt{\frac{1}{N}\sum_i f_i (x_i - \bar{x})^2}$$

Where $\bar{x}$ is the arithmetic mean of the distribution.

The step of squaring the deviations (xi - $\bar{x}$) overcomes the drawback of ignoring the signs in mean deviation. Standard deviation is also suitable for further mathematical treatment.

Thus we see that standard deviation satisfies almost all the properties laid down for an ideal measure of dispersion except for the general nature of extracting the square root which is not readily comprehensible for a non-mathematical person. It may also be pointed out that standard deviation gives greater weight to extreme values and as such has not found favor with economists or businessmen who are more interested in the results of the modal class'. Taking into Consideration the pros and cons and also the wide applications of standard deviation in statistical theory, we may regard standard deviation as the best and the most powerful measure of dispersion.

The square of standard deviation is called the variance and is given by

$$\sigma^2 = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^2$$

(d) Define the coefficients of skewness, kurtosis and variation with one example of each.

Ans: - Coefficients of Skewness: Literally, skewness means 'lack of symmetry·. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data, A distribution is said to be skewed if

i) Mean, median and mode fall at different points

i.e., Mean ≠Median≠ Mode,

ii) Quartiles are not equidistant from median, and

iii) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other

Measures of Skewness: Various measures of skewness are

(1) $S_k = M - M_d$

(2) $S_k = M - M_o$

Where M is mean, M_d, the median and M_o the mode of the distribution.

(3) $S_k = (Q_3 - M_d) - (M_d - Q_1)$

These are the absolute measures of skewness. As in dispersion, for comparing two series we do not calculate these absolute measures but we calculate the relative measures called the co-efficient of skewness.

Kurtosis : If we know the measures of central tendency, dispersion and skewness, we still cannot form a complete idea about the distribution' as will be clear from the following figure in which all the three curves A, Band Care symmetrical about the mean om' and have the same range.

In addition to these measures we should know one more measure which Prof. Karl Pearson calls as the 'Convexity of curve or Kurtosis. Kurtosis enables us to have an idea about the flatness or peakedness of the curve, It is measured by the co-efficient $\beta_2$or its derivation r2 given by

$$\beta_2 = \mu_4/\mu_2^2, \quad \gamma_2 = \beta_2 - 3$$

Q2. (a) Define regression. State the properties of regression coefficient. Derive an equation for the line of regression of Y on X.

Ans:- Regression: Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.

In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the values or is used for prediction. is called independent variable. In regression analysis independent variable is also known as regressor or predictor or explanatory variable while the dependent variable is also known as regressed or explained variable.

Properties of Regression of Coefficient:-

(a) Correlation coefficient is the geometric mean between the regression coefficients.

Proof:- Multiplying we get

$$b_{XY} \times b_{YX} = r\frac{\sigma_X}{\sigma_Y} \times r\frac{\sigma_Y}{\sigma_X} = r^2$$

$$r = \pm\sqrt{b_{XY} \times b_{YX}}$$

It may be noted that the sign of correlation coefficient is the same as that of regression coefficients, since the sign of each depends upon the co-variance term μ_{11}. Thus if the regression coefficients are positive, 'r' is positive and if the regression coefficients are negative 'r' is negative.

From above equation, we have

$$r = \pm\sqrt{b_{XY} \times b_{YX}}$$

The sign to be taken before the square root is that of regression coefficients.

(b) If one of the regression coefficients is greater than unity, the other must be less than unity

Proof:- Let one of the regression coefficients (say) b_{YX} be greater than unity, then we have to show that $b_{YX} < 1$.

Now $b_{YX} > 1 \Rightarrow 1/b_{YX} < 1$

Also $r^2 <= 1 \Rightarrow b_{YX} \cdot b_{XY} <= 1$

Hence $b_{YX} < = 1/b_{YX} < 1$

(c) Arithmetic mean of the regression coefficients is greater than the correlation coefficient r, provided r>0.

Proof:- We have to prove that ½ ($b_{YX} + b_{YX}$) >= r

$$\frac{1}{2}\left(r\frac{\sigma_Y}{\sigma_X} + r\frac{\sigma_X}{\sigma_Y}\right) \geq r \quad \text{or} \quad \frac{\sigma_Y}{\sigma_X} + \frac{\sigma_X}{\sigma_Y} \geq 2$$

Or $\sigma_Y^2 + \sigma_X^2 - 2\sigma_X\sigma_Y \geq 0 \quad i.e., \quad (\sigma_Y - \sigma_X)^2 \geq 0$

Which is always true, since the square of a real quality is >= 0.

(d) Regression coefficients are independent of the change of the origin but not scale.

Proof:- Let $U = \frac{X-a}{h}$, $V = \frac{Y-b}{k}$ $\Rightarrow X = a + hU, Y = b + kV,$

Where a,b,h (>0) and k (>0) are constants.

Then Cov (X, Y) = hk Cov (U,V), $\sigma_X^2 = h^2\ \sigma_U^2$ and $\sigma_Y^2 = k^2\ \sigma_V^2$

$$b_{YX} = \frac{\mu_{11}}{\sigma_X^2} = \frac{hk\ \text{cov}\ (U, V)}{h^2\sigma_U^2}$$
$$= \frac{k}{h} \cdot \frac{\text{cov}\ (U, V)}{\sigma_U^2} = \frac{k}{h} b_{VU}$$

Similarly, we can prove that

$$b_{XY} = (h/k)\ b_{UV}$$

Line of Regression: If the variables in a bivariate distribution are related. we will find that the points in the scatter diagram will cluster round some curve called the "curve of regression". If the curve is a straight line, it is called the line of regression and there is said to be linear regression' between the variables, otherwise regression is said to be curvilinear.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of "best fit" and is obtained by the principles of least squares.

Let us suppose that in the bivariate distribution (xi,yi); i=1,2,….,n; Y is dependent variable and X is independent variable. Let the line of regression of Y on X be Y = a+bX.

According to the principle of least square, the normal equations for estimating a and b are

$\sum_{i=1}^{n} y_i = na + b\ \sum_{i=1}^{n} x_i$ ……(1)

And $\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b\ \sum_{i=1}^{n} x_i^2$ ……(2)

From (1) on dividing by n, we get

$$\bar{y} = a + b\bar{x}$$

Thus the line of regression Y on X passes through the point $(\bar{x}, \bar{y})$

Now

$$\mu_{11} = COV(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\bar{y}$$

……..(3)

Also $\sigma_{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x^2} \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma_{x^2} + \bar{x}^2$ ……..(3a)

Dividing (2) by n and using (3) and (3a) we get

$$\mu_{11} + \bar{x}\,\bar{y} = a\bar{x} + b(\sigma_X^2 + \bar{x}^2)$$ …….. (4)

Multiplying by $\bar{x}$ and then subtracting from (4) we get

$$\mu_{11} = b\sigma_X^2 \quad \Rightarrow \quad b = \frac{\mu_{11}}{\sigma_X^2}$$ ……..(5)

Since ‘b’ is the slope of the line of regression of Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y})$, its equation is

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x})$$

$$\Rightarrow \qquad Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

Starting with the equation X= A +BY and proceeding similarly or by simply interchanging the variables X and Y in above equations, the equation of the line of regression of X on Y becomes

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y})$$

$$\Rightarrow \quad X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

(b) Explain the following measures of dispersion:

(i) Range

(ii) Quartile deviation

(iii) Mean deviation

(i) Standard deviation.

Ans:- **(i) Range :** The range is the difference between two extreme observations, or the distribution. If A and B are the greatest and smallest observations respectively in a distribution, then its range is A-B.
Range is the simples I but a crude measure of dispersion. Since it is based on two extreme observations which themselves are subject to change fluctuations, it is not at all a reliable measure of dispersion.

(ii) Quartile Deviation : Quartile deviation or semi – interquartile range Q is given by

$$\mathbf{Q = \tfrac{1}{2}(Q_3 - Q_1),}$$

Where Q1 and Q3 are the first and third quartile of the distribution respectively.

Quartile deviation is definitely a better measure than the range as it makes use of 50% of the data. But since it ignores the other 50% of the data, it cannot be regarded as a reliable measure.

(iii)Mean Deviation : If xi| fi, i=1,2,….,n is the frequency distribution, then mean deviation from the average A, (usually mean, median or mode) is given by

$$\text{Mean deviation} = \frac{1}{N} \sum_i f_i |x_i - A|, \ \Sigma f_i = N$$

Where |xi – A| represents the modulus or the obsolute value of the deviation (xi – A), when the –ve signed is ignored.
Since mean deviation is based on all the observations, it is a better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviations (xi – A) create artificiality and renders it useless for further mathematical treatment.

(iv)Standard Deviation : Standard deviation, usually denoted by the Greek letter small sigma (σ), is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. For the frequency distribution xi|fi, i= 1,2,…,n

$$\sigma = \sqrt{\frac{1}{N} \sum_i f_i (x_i - \bar{x})^2}$$

Where $\bar{x}$ is the arithmetic mean of the distribution.

The step of squaring the deviations (xi - $\bar{x}$) overcomes the drawback of ignoring the signs in mean deviation. Standard deviation is also suitable for further mathematical treatment.

Thus we see that standard deviation satisfies almost all the properties laid down for an ideal measure of dispersion except for the general nature of extracting the square root which is not readily comprehensible for a non-mathematical person. It may also be pointed out that standard deviation gives greater weight to extreme values and as such has not found favor with economists or businessmen who are more interested in the results of the modal class'. Taking into Consideration the pros and cons and also the wide applications of standard deviation in statistical theory, we may regard standard deviation as the best and the most powerful measure of dispersion.

The square of standard deviation is called the variance and is given by

$$\sigma^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2$$

(c) Explain skewness and kurtosis in detail with the help of diagram.

Ans:- Skewness:- Skewness means 'lack of symmetry'. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. A distribution is said to be skewed if

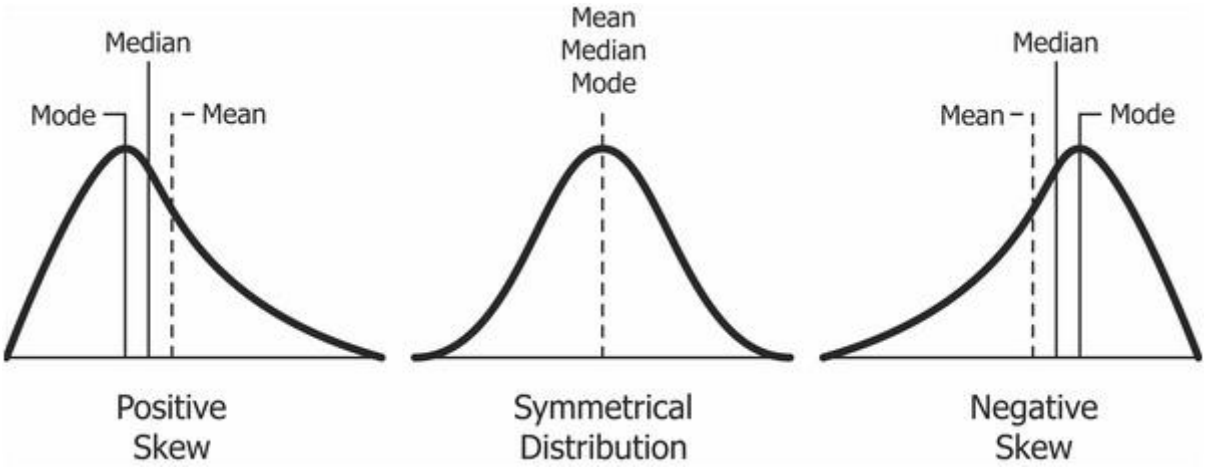
(i) Mean, median and mode fall at different points,

i.e. Mean ≠ Median ≠ Mode

(ii) Quartile are not equidistant from median, and

(iii) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

The figure of the symmetrical, positively skewed and negatively skewed distribution are given below:

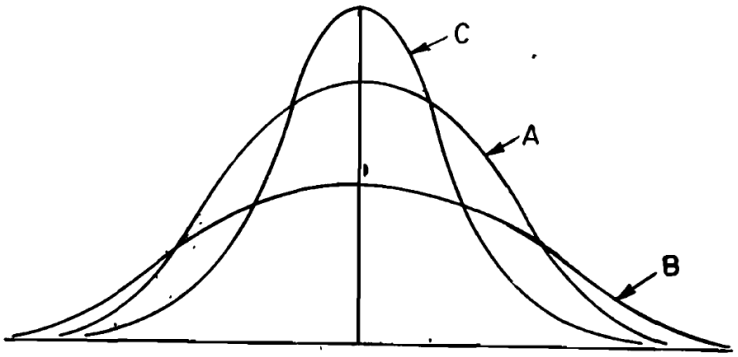


Kurtosis : If we know the measures of central tendency, dispersion and skewness, we still cannot form a complete idea about the distribution' as will be clear from the following figure in which all the three curves A, Band Care symmetrical about the mean om' and have the same range.

In addition to these measures we should know one more measure which Prof. Karl Pearson calls as the 'Convexity of curve or Kurtosis. Kurtosis enables us to

have an idea about the flatness or peakedness of the curve, It is measured by the co-efficient $\beta_2$or its derivation r2 given by

$\beta_2 = \mu_4/\mu_2^2$, $\gamma_2 = \beta_2 - 3$



Curve of the type 'A' which is neither flat nor peaked is called the normal curve or mesokurtic curve and for such a curve β_2 = 3, i.e. $\gamma_2 = 0$. Curve of the type 'B' which is flatter than the normal curve is known as platykurtic and for such a curve β_2 < 3, i.e. $\gamma_2 < 0$. Curve of the type 'C' which is more peaked than the normal curve is called leptokurtic and for such a curve β_2 > 3, i.e. $\gamma_2 > 0$

(d) Define correlation and coefficient of correlation. Explain scatter diagram in detail.

Ans:- Correlation :- Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

Correlation Coefficient: A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. In positively correlated variables, the value increases or decreases in tandem. In negatively correlated variables, the value of one increases as the value of the other decreases.

Correlation coefficients are expressed as values between +1 and -1. A coefficient of +1 indicates a perfect positive correlation: A change in the value of one variable will predict a change in the same direction in the second variable. A coefficient of -1 indicates a perfect negative correlation: A change in the value of one variable predicts a change in the opposite direction in the second variable. Lesser degrees of correlation are expressed as non-zero decimals. A coefficient of zero indicates there is no discernable relationship between fluctuations of the variables.

Scatter Diagram:- It is the simplest way of the diagrammatic representation of bivariate data. Thus for the bivariate distribution (xi, yi); I = 1,2,…n, if the values of the variable X and Y be plotted along the x – axis and y – axis respectively in the xy plane, the diagram of dots so obtained is known as scatter diagram. From the scatter diagram, we can form a fairly good, though vague, idea whether the variable are correlated or not, e.g., if the points are very dense, i.e., very close to each other, we should expect a fairly good amount of correlation between the variables and if the points are widely scattered, a poor correlation is expected.

UNIT – III

Q 3

(A) Explain:

(i) Mathematical law of addition of probabilities

(ii) Conditional Probability

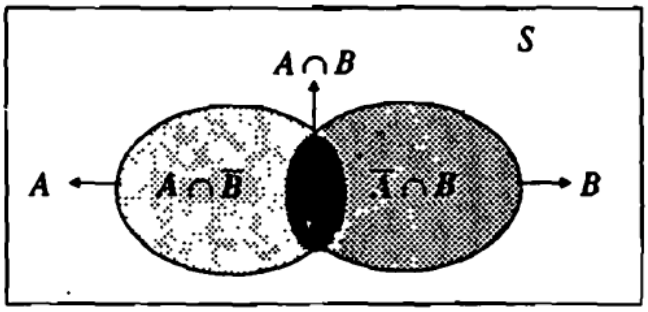
Ans: -

(i) Mathematical law of addition of probability

Theorem: If A and B are any two events (subsets of sample space S) and are not disjoint, then

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

Proof:



From the Venn diagram, we have

$A \cup B = A \cup (\bar{A} \cap B)$

Where A and $\bar{A} \cap B$ are mutually disjoint.

$\therefore \quad P(A \cup B) = P[A \cup (\bar{A} \cap B)]$

$= P(A) + P(\bar{A} \cap B)$

$= P(A) + P(B) - P(A \cap B)$ [From theorem]

(ii) Conditional Probability

The probability P(A) of an event A represents the likelihood that a random experiments will result in an outcome in the set A relative to the sample space of S of the random experiments. However, quite often, while evaluating some event probability, we already have some information that the outcome of the random experiments. For example, if we have prior information that the outcome of the random experiments must be in a set of B of S, then this information must be used to re-appraise the likelihood that the outcome will also be in B. This re-appraised probability is denoted by P(A|B) and is read as the conditional probability of the event A, give that the event B has already happened.

(b) What is the probability that at least two out of n people have the same birthday? Assume 365 days in a year and that all days are equally likely. (S – 14)

Ans: Since the birthday of any person can fall on any one of the 365 days, the exhaustive number of cases for the birthdays of n persons in 365"

If the birthdays of all n persons fall on different days, then the number of favourable cases is :

365 (365-1) (365-2) ... [365-(n-1)], because in this case the birthday of the first person can fall on any one of 365 days, the birthday of the second person cal fall on any one of the remaining 364 days, and so on. Hence, the probability (p) that birthdays of all the n persons are different is given by:

$$\rho = \frac{365(365-1)(365-2)\ldots\{365-(n-1)\}}{365^n}$$

$$\rho = \left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\left(1-\frac{3}{365}\right)\ldots\left(1-\frac{n-1}{365}\right)$$

Hence, the required probability that at least two persons have same birthday is :

$$1-\rho = 1-\left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\left(1-\frac{3}{365}\right)\ldots\left(1-\frac{n-1}{365}\right)$$

(C) State and prove Baye's Theorem (S – 14)

Ans: - Baye's Theorem: - If $E_1,E_2,E_3,...,E_n$ are mutually disjoint events with $P(E_i) \neq 0$, (i=1 ,2, ..., n), then for any arbitrary event A which is a subset of $\bigcup_{i=1}^{n} E_i$ such that P(A) > 0 , we have

$$P(E_i \mid A) = \frac{P(E_i)\,P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\,P(A \mid E_i)}, \quad i = 1, 2, ..., n.$$

Proof: -Since $A \subset \bigcup_{i=1}^{n} E_i$, we have, $A \cap (\bigcup_{i=1}^{n} E_i) = \bigcup_{i=1}^{n} (A \cap E_i)$

[By distributive law]

Since $(A \cap E_i) \subset E_i$, (i=1,2,...,n) are mutually disjoint events, we have by addition theorem of probability:

$$P(A) = P\left[\bigcup_{i=1}^{n} (A \cap E_i)\right] = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(E_i)\,P(A \mid E_i),$$

By multiplication theorem of probability

Also we have

$$P(A \cap E_i) = P(A)\,P(E_i \mid A)$$

$$P(E_i \mid A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(E_i)\,P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\,P(A \mid E_i)}$$

(D) **Explain Geometric probability**

Ans: Geometric probability: It was pointed out that the classical. Definition of probability fails if the total number of outcomes of an experiment is infinite. Thus, for example. if we are interested in finding the probability that a point selected at random in a given region will lie in a specified part of it, the classical definition of probability is modified and extended to what is called **geometrical probability** or probability in continuum, In this case, the general 'expression for probability 'p' is given by

$$p = \frac{\textit{Measure of specified part of the region}}{\textit{Measure of the whole region}}$$

Q3. (a) A box contains 4 red, 2 white, and 3 black balls. A person draws one ball from the box at random. Find the probability that among the balls drawn there is at least one ball of each colour.

Ans: The required event E that in a draw of one ball from the box at random there is at least one ball of each colour, can materialize in the following mutually disjoint ways

(i) 1 red
(ii) 1 white
(iii) 1 black

Hence by the addition theorem of probability, the required probability is given by

P(E) = P(i) + P(ii) + P(iii)

$$= \frac{{}^4C_1 \times {}^2C_1 \times {}^3C_1}{{}^9C_1} + \frac{{}^4C_1 \times {}^2C_1 \times {}^3C_1}{{}^9C_1} + \frac{{}^4C_1 \times {}^2C_1 \times {}^3C_1}{{}^9C_1}$$

$$= \frac{1}{{}^9C_1}[4 \times 2 \times 3 + 4 \times 2 \times 3 + 4 \times 2 \times 3]$$

$$= \frac{1}{9 \times 8}[24 + 24 + 24] = 1$$

(b) State and prove Baye's Theorem

Ans: - Baye's Theorem: If E1,E2, …… , En are mutually disjoint with $p(E_i) \neq 0$, (i=1,2,…..,n) then for any orbitory event A which is a subset of $\coprod_{i=1}^{n} E_i$, such that P(A)>0,

We have

$$P(E_i \mid A) = \frac{P(E_i)\,P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\,P(A \mid E_i)}, \quad i = 1, 2, \ldots, n. \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Proof: Since $A \subset \bigcup_{i=1}^{n}(E_i)$, we have

$$A = A \cap (\bigcup_{i=1}^{n}(E_i) = \bigcup_{i=1}^{n}(A \cap E_i)$$ [By distributive law]

Since $(A \cap E_i) \subset E_i$ (i=1,2,…..,n) are mutually disjoint events, we have given by addition theorem of probability

$$P(A) = P[\bigcup_{i=1}^{n}(A \cap E_i)] = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(E_i)\,P(A \mid E_i), \qquad \ldots(*)$$

By compound theorem of probability.

Also we have

$$P(A \cap E_i) = P(A)\,P(E_i \mid A)$$

$$P(E_i \mid A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(E_i)\,P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\,P(A \mid E_i)} \qquad [\text{From } (*)]$$

Hence proved.

(c) What are the independent events? Explain with its multiplication theorem of probability.

Ans: - Independent Events : - An event B is said to be independent of event A, if the conditional probability of B given A i.e. P(B|A) is equal to the unconditional probability of B i.e., if

$$P(B|A) = P(B)$$

Since

$$P(A \cap B) = P(B \mid A)\,P(A) = P(A \mid B)\,P(B)$$

And since P(B|A) = P(B) when B is independent of A, we must have P(A|B) = P(A) or it follow that A is also independent of B. hence the events A and B are independent if and only if

$$P(A \cap B) = P(A)\,P(B)$$

Multiplication Theorem of Probability:

For two events A and B

$$P(A \cup B) = P(A).P(B|A), P(A) > 0$$
$$= P(B).P(A|B), P(B) > 0$$

Where P(B|A) represents the conditional probability of occurrence of B when the event A has already happened and P(A|B) is the conditional probability of happening of A, given that B has already happened.

Proof:

$$P(A) = \frac{n(A)}{n(S)} \; ; \; P(B) = \frac{n(B)}{n(S)} \text{ and } P(A \cap B) = \frac{n(A \cap B)}{n(S)} \qquad (*)$$

For the condition event A|B, the favourable outcomes must be one of the sample points of B, i.e. for the event of B, i.e., for the event A|B, the sample space is B and out of the n(B) sampl;e points, n(A$\cap B$) pertain to the occurrence of the event A. Hence

$$P(A \mid B) = \frac{n(A \cap B)}{n(B)}$$

Rewriting (*), we get

$$P(A \cap B) = \frac{n(B)}{n(S)} \quad \frac{n(A \cap B)}{n(B)} = P(B) \; . \; P(A \mid B)$$

Similarly we can prove:

$$P(A \cap B) = \frac{n(A)}{n(S)} \; . \; \frac{n(A \cap B)}{n(A)} = P(A) \; . \; P(B \mid A)$$

Multiplication Law of probability for Independent Events:

If A and B are independent then

P(A|B) =P(A) and P(B|A) = P(B)

Hence it gives:

$P(A \cap B) = P(A)P(B)$

Provided A and B are independent.

(d) Two groups of subjects contain respectively 3 female and 2 male students, and 3 female and 4 male students. Two students are selected at random from each group. Find the chance for one female and one male student selection.

Ans:- The required event of getting one female and 1 male among the two selected students can materialize in the following two mutually disjoint cases:

Group No.	I	II
(i)	Female	Male
(ii)	Male	Female

Hence by addition theorem of probability

Required probability = P(i) + P(ii)

`Since the probability of selecting a female from first group is 3/5, of selecting a male from the second group is 4/7 and since these two events of selecting two students from two group are independent of each other, by compound probability theorem, we have

$$P(i) = \frac{3}{5} \times \frac{4}{7} = \frac{12}{35}$$

$$P(ii) = \frac{1}{3} \times \frac{1}{4} = \frac{1}{12}$$

Substituting the value of P(i) and P(ii), we get

Required probability = $= \frac{12}{35} + \frac{1}{12} = \frac{179}{420} = 0.42$

Q3(a). Explain the following terms with example:

(i) Exhaustive events

(ii) Mutually exclusive events

(iii) Equally likely events

(iv) Independent events.

Ans:- (i) Exaustive Events:- The total number of possible outcomes in any trial is known as exhaustive events. For example : (a) In tossing of a coin there are two exhaustive cases, viz., head and tail.

(b) In throwing of a die, there are six exhaustive cases since any one of the 6 faces 1,2, …6 may come uppermost.

(ii) Mutually exclusive events: - Events are said to be mutually exclusive or incompatible if the happening of any one of them precludes the happening of all the others, i.e. no two or more of them can happen simultaneously in the same trial. For example: (a) In throwing a die all the 6 faces numbered 1 to 6 are mutually exclusive since if any one of these faces comes, the possibilities of others, in the same trial, is rules out.

(iii) Equally likely events :- Outcomes of a trial are set to be equally likely if taking into consideration all the relevant evidences, there is no reason to expect one in preference to the others. For example:- (a) In tossing an unbiased or uniform coin, head or tail are equally likely events.

(iv)**Independent events:-** Several events are said to be independent if the happening of an event is not affected by the supplementary knowledge concerning the occurrence of any number of the remaining events. For example:- (a) In tossing an unbiased coin the event of getting a head in the first toss is independent of getting a head in the second, third and subsequent throws.

(b) State and Prove Baye's Theorem.

Ans:- Baye's Theorem: - If $E_1, E_2, E_3, ..., E_n$ are mutually disjoint events with $P(E_i) \neq 0$, (i=1 ,2, ..., n), then for any arbitrary event A which is a subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$, we have

$$P(E_i \mid A) = \frac{P(E_i)\,P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\,P(A \mid E_i)}, \quad i = 1, 2, \ldots, n.$$

Proof: -Since $A \subset \bigcup_{i=1}^{n} E_i$, we have, $A \cap (\bigcup_{i=1}^{n} E_i) = \bigcup_{i=1}^{n} (A \cap E_i)$

[By distributive law]

Since $(A \cap E_i) \subset E_i$, (i=1,2,...,n) are mutually disjoint events, we have by addition theorem of probability:

$$P(A) = P\left[\bigcup_{i=1}^{n} (A \cap E_i)\right] = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(E_i)\,P(A \mid E_i),$$

By multiplication theorem of probability

Also we have

$$P(A \cap E_i) = P(A)\,P(E_i \mid A)$$

$$P(E_i \mid A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(E_i)\,P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\,P(A \mid E_i)}$$

(c) State and prove the :
(i) Addition law of probability for Mutually exclusive.
(ii) Addition law of probability for Not Mutually exclusive.

Ans:- (i) Addition law of probability for Mutually exclusive:-

(d) The probability of 'X', 'Y' and 'Z' becoming managers are 4/9, 2/9 and 1/3 respectively. The probabilities that the bonus scheme will be introduced if 'X', 'Y' and 'Z' becomes manager are 3/10, 1/2 and 3/5 respectively.
(i) What is the probability that bonus scheme will be introduced and
(ii) If the bonus scheme has been introduced, what is the probability the manager appointed was 'X' ?

Ans:- Let B1, B2, B3 be the evens that respectively X, Y, Z becomes manager, and A the event that Bonus scheme is introduced. We have

$P(B_1) = 4/9$, $P(B_2) = 2/9$, $P(B_3) = 1/3$,

$P(A|B_1) = 3/10$, $P(A|B_2) = ½$, $P(A|B_3) = 4/5$

(i) We have, $A = \bigcup_{1=1}^{3}(A \cap B_i)$, where $A \cap B_i$, i= 1,2,3 are mutually exclusive, thus

$$P(A) = \sum_{i=1}^{3} P(B_i)P(A|B_i) = \frac{4}{9} \times \frac{3}{10} + \frac{2}{9} \times \frac{1}{2} + \frac{1}{3} \times \frac{4}{5} = \frac{2}{15} + \frac{1}{9} + \frac{4}{15} = \frac{23}{45}$$

(ii) $B_1 \cup B_2$ is the event that manager appointed was X or Y, also $B_1 \cap B_2 = \emptyset$. Thus,

$$P(B_1 \cup B_2|A) = P(B_1|A) + P(B_2|A)$$

$$= \frac{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}{P(A)}$$

$$= \frac{\frac{2}{15} + \frac{1}{9}}{\frac{23}{45}} = \frac{22}{90} \times \frac{45}{23} = \frac{11}{23}$$

UNIT – IV

(B) Prove that: If F is the density function of the random variable X and if a<b, then P(a ≤ X ≤ b) = F(b) – F(a). (S – 14)

Ans: - The event ' a < X ≤ b' and X≤ a are disjoint and their union is the event 'X≤b'. Hence by addition theorem of probability:

P (a < X ≤b) + P (X≤ a) = P (X ≤ b)

$$\Rightarrow P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$

$$\Rightarrow P(a < X \leq b) = F(b) - F(a)$$

(c) A continuous random variable X has a probability density function $f(x) = 4x^2$, $0 \leq x \leq 1$. Find a and b such that

(i) P(x ≤ a) = P(x>a)

(ii) P (x >b) = 0.05

Ans: - (i) Since P (X ≤ a) = P(X>a), each must be equal to ½, because total probability is always unity.

$$\therefore \quad P(X \leq a) = \frac{1}{2} \quad \Rightarrow \quad \int_0^a f(x)ds = \frac{1}{2}$$

$$\Rightarrow \quad 4\int_0^a x^2\,dx = \frac{1}{2} \qquad \Rightarrow \quad 4\left|\frac{x^3}{3}\right|_0^a = \frac{1}{2} \qquad \Rightarrow a = \left(\frac{1}{2}\right)^{\frac{1}{4}}$$

$$(ii) \quad P(X > b) = 0.05 \qquad \Rightarrow \int_b^1 f(x)dx = 0.05$$

$$\Rightarrow \quad 4\left|\frac{x^3}{3}\right|_b^1 = \frac{1}{20} \qquad \Rightarrow \quad 1 - b^3 = \frac{1}{20} \quad \Rightarrow \quad b = \left(\frac{19}{20}\right)^{\frac{1}{4}}$$

(d) Define Mathematical expectation. Hence state and prove addition and multiplication Theorem of Expectation. (S -14)

Ans: The expected value of a discrete random variable is a weighted average of all possible values of the random variable, where the weights are the probabilities associated with the corresponding values. The mathematical expression for computing the expected value of a discrete random variable X with probability mass function (p.m.f.) f(x) is given below:

$$E(X) = \sum_{x} xf(x), \quad (for\ discrete\ r.v.)$$

The mathematical expression for computing the expected values of a continuous random variable X with probability density function (p.d.f.)f(x) is, however, as follows:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx, \qquad (for\ continuous\ r.v.)$$

Addition Theorem of Expectation:

If X and Y are random variable, then E(X+Y) = E(X) + E(Y)

Provided all the expectations exist.

Prof: - Let X and Y be continuous r.v. with joint p.d.f. f_{XY} (x,y) and marginal p.d.f's f_x(x) and f_y (y) respectively. Then by def.,

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \ldots (1) \qquad \text{and} \quad E(Y) = \int_{-\infty}^{\infty} yf(y)dy \ldots (2)$$

$$E(X+Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)f_{XY}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf_{XY}(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf_{XY}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f_{XY}(x,y)dy\right]dx + \int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f_{XY}(x,y)dx\right]dy$$

$$= \int_{-\infty}^{\infty} xf(x)dx + \int_{-\infty}^{\infty} yf(y)dy$$

= E(X) + E(Y)

Multiplication Theorem of Expectations

If X and Y are independent random variable, then E(XY) = E(X) . E(Y)

Proof: - Proceeding as in property 1 we have

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{XY}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x)\, f_Y(y)dxdy$$

$$= \int_{-\infty}^{\infty} x f_X(x)\, dx \int_{-\infty}^{\infty} y\, f_Y(y)dy$$

= E(X) E(Y) , provided X and Y are independent.

Q4. (a) Explain variance and covariance of random variable, find E(x) and E(x^2) for random variable x with following probability distribution.

X:	**-3**	**6**	**9**
P(X=x)	**1/6**	**½**	**1/3**

Ans: - **Covariance:-** If X and Y are two random variables, then covariance between them is defined as

Cov (X,Y) = E[{X – E(X) } {Y – E(Y)}]
= E[XY – X E(Y) – YE(X) + E(X) E(Y)]
= E(XY) – E(Y) E(X) – E(X) E(Y) + E(X) E(Y)

If X and Y are independent then E(XY) = E(X) E(Y) and hence in this case

Cov (X,Y) = E(X) E(Y) – E(X) E(Y) = 0

find E(x) and E(x^2) for random variable x with following probability distribution.

X:	**-3**	**6**	**9**

P(X=x)	1/6	½	1/3

$E(X) = \sum x.p(x)$

$= (-3) \times 1/6 + 6 \times ½ + 9 \times 1/3 = 11/2$

$E(X^2) = \sum x^2.p(x)$

$= 9 \times 1/6 + 36 \times ½ + 81 \times 1/3 = 93/2$

(b) Define Mathematical Expectation. Derive multiplication theorem of expectation.

Ans: - **Mathematical Expectation:-** Let X be random variable (r.v.)with p.d.f. (p.m.f.). Then its mathematical expectation, denoted by E(X) is given :

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx, \quad (\text{for continuous } r.v.)$$

$$= \sum_{x} x\, f(x), \quad (\text{for discrete } r.v.)$$

Provided the right-hand integral or series is absolute convergent, i.e., provided

$$\int_{-\infty}^{\infty} |xf(x)|dx = \int_{-\infty}^{\infty} |x|f(x)dx < \infty$$

Or

$$\sum_x |xf(x)| = \sum_x |x|f(x) < \infty$$

Multiplication Theorem of Expectation

Theorem: If X and Y are independent random variable, then

E(XY) = E(X)E(Y)

Proof: As given

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{XY}(x,y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x)\, f_Y(y)\, dx\, dy$$

[Since X and Y are independent]

$$= \int_{-\infty}^{\infty} x\, f_X(x)\, dx \int_{-\infty}^{\infty} y\, f_Y(y)\, dy$$

$$= E(X).E(Y), \quad \textit{[Using (6·11) and (6·12)]}$$

Provided X and Y are independent.

(c) Define probability mass function. A random variable x has the following probability function.

X	0	1	2	3	4	5	6	7
P(x)	0	K	2K	2K	3K	K^2	$2K^2$	$7K^2+K$

Determine K and F(x)

Ans:- Since $\sum_{x=0}^{7} p(x) = 1$, we have

$K+2k+2k+3k+k^2+2k^2+7k^2+k = 1$

$10k^2 + 9k - 1 = 0$

$(10k - 1)(k + 1) = 0$

$K = 1/10$ (since k=-1 is rejected, since probability cannot be negative)

For calculating F(x)

X	F(x) = P(X <= x)
0	0
1	K = 1/10
2	3k = 3/10
3	5k = 5/10
4	8k = 4/5
5	$8k + k^2 = 81/100$
6	$8k + 3k^2 = 83/100$
7	$9k + 10k^2 = 1$

(d) Define : -

(i) Discrete random variable

(ii) Continuous random variable

(iii) Moment generation function

(iv) Probability distribution function

Ans: -

(ii) **Discrete random variable :** - If a random variable takes at most a countable number of values, it is called a discrete random variable. In other words, a real valued function defined on a discrete sample space is called a discrete random variable.

(iii) **Continuous Random variable :** - A random variable X is said to be continuous if it can take all possible values between certain limits. In other words, a random variable is said to be continuous when its different values cannot be put in 1 – 1 correspondence with a set of positive integers

A continuous random variable is a random variable that can be measured to any desired degree of accuracy. Examples of continuous random variables are age, height, weight etc.

(iv) Moment generating function :- The moment generating function (m.g.f.) of a random variable X (about origin) having the probability function f(x) is given by

$$\begin{cases} M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx, \\ \qquad\qquad \text{(for continuous probabiliity distribution)} \\ = \sum_x e^{tx} f(x), \\ \qquad\qquad \text{(for discrete probability distribution)} \end{cases}$$

the integration or summation being extended to the entire range of x, t being the real parameter and it is being assumed that the right-hand side of above equation is absolutely convergent for some positive number h such that –h < t < h. thus

$$M_X(t) = E(e^{tX}) = E\left[1 + tX + \frac{t^2X^2}{2!} + \ldots + \frac{t^rX^r}{r!} + \ldots\right]$$

$$= 1 + t\,E(X) + \frac{t^2}{2!}E(X^2) + \ldots + \frac{t^r}{r!}E(X^r) + \ldots$$

$$= 1 + t\,\mu_1' + \frac{t^2}{2!}\mu_2' + \ldots + \frac{t^r}{r!}\mu_r' + \ldots$$

where $\mu_r' = E(X^r) = \int x^r f(x)\,dx$, *for continuous distribution*

$= \Sigma\, x^r p(x)$, *for discrete distribution,*

is the rth moment of X about origin. Thus we see that the coefficient of $\frac{t^r}{r!}$ in Mx(t) gives u_r' (above origin). Since Mx(t) generate moments, it is known as moment generating function.

Q4. (a) Define discrete random variable. Comment on probability mass function and probability density function. Define Mathematical expectation.

Ans:- Discrete Random Variable:- If a random variable takes at most a countable number of values, it is called a discrete random variable. In other words, a real valued function defined on a discrete sample space is called a discrete random variable.

Probability Mass Function:- Suppose X is one-dimensional discrete random variable taking at most a countably infinite number of values x1,x2,.. with each possible outcome xi, we associate a number pi = P(X = xi) = p(xi), called the probability of xi. The numbers p(xi); I = 1,2,… must satisfy the following condition

(i) p(xi) >= 0 for all i

(ii) $\sum_{i=1}^{\infty} p(x_i) = 1$

This function p is called the probability mass function of the random variable X.

Probability Density Function:- Probability density function f_x (x) of the random variable X is defined as:

$$f_x(x) = \lim_{\delta x \to 0} \frac{P(x \le X \le x + \delta x)}{\delta x}$$

The probability for a variate value to lie in the interval dx is f(x) dx and hence the probability for a vairate value to fall in the finite interval [α,β] is :

$$P(\alpha \le X \le \beta) = \int_{\alpha}^{\beta} f(x)\,dx$$

Which represents the area between the curve y = f(x), x – axis and the ordinates at x = α and x= β. Further since total probability is unity, we have

$\int_a^b f(x)dx = 1$, where [a,b] is the range of the random variable X. The range of the variable may be finite or infinite.

The probability density function of random variable X usually denoted by $f_x(x)$ or simply f(x) has the following obvious properties

(i) F(x) >= 0, $-\infty < x < \infty$

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

(iii) The probability P(E) given by

$$P(E) = \int_E f(x)\,dx$$

Is well defined for any event E.

Mathematical expectation :- Let X be a random variable (r. v.) with p.d.f. f(x). Then its mathematical expectation, denoted by E(X) is given by:

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)dx, \quad \text{(for continuous r.v)}$$

$$= \sum x\, f(x) \qquad \text{(for discrete r. v.)}$$

UNIT – V

Q5 (A) Explain Binomial distribution with appropriate example. (S – 14)

Ans: - **Binomial distribution: L**et a random experiment be performed repeatedly and let the occurrence of an event in a trial be called a success and its non-occurrence a failure. Consider a set of n independent Bernoullian trials (b being finiteI in which the probability 'P' of success in any trial constant for each trial, then q=1-p, is the probability of failure in any trial.

The probability of x successes and consequently (n-x) failure in n independent trials, in a specified order (say) SSFSFFFS...FSF (where S represents success and F represents failure) is given by the compound probability theorem by the expression:

P(SSFSFFFS...FSF) = P(S)P(S)P(F)P(S)P(F)P(F)P(F)P(S) X... XP(F)P(S)P(F)

= p.p.q.p.q.q.q.p.... q.p.q

= p. p. p. ... p . q.q.q. .....q = $p^x q^{n-x}$

{x factors} {(n-x) factors}

But x successes in n trials can occurs in (n/x) ways and the probability for each of these ways is same is, viz., $p^x q^{n-x}$. Hence the probability by the expression (n/x) $p^x q^{n-x}$

The probability distribution of the number of successes, so obtained is called the Binomial probability distribution, for the obvious reason that the probabilities of 0,1,2, ...,n successes, viz., $q^n, \binom{n}{1}, q^{n-1}p, \binom{n}{2} q^{n-2}p^2, \ldots, p^n$ are the successive terms of the binomial expansion $(q+p)^n$

Definition: A random variable X is said to follow binomial distribution if it assumes only non-negative values and ts probability mass function is given by:

$$P(X = x) = p(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}; x = 0,1,2, \ldots, n \; ; q = 1 - p \\ 0 \qquad\qquad , \text{otherwise} \end{cases}$$

The two independent constant n and p in the distribution are known as the parameters of the distribution. 'n' is also sometimes, known as the degree of the binomial distribution.

Ex. Ten coins are thrown simultaneously. Find the probability of getting at least seven heads.

Son:- p = Probability of getting head = ½

q = Probability of not getting a head =1/2

The probability of getting x heads in a random thrown of 10 coins is :

$$p(x) = \binom{10}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{10-x} = \binom{10}{x}\left(\frac{1}{2}\right)^{10} ; x = 0,10,2,\dots,3$$

Probability of getting at least seven heads is given by:

$(P \geq 7) = p(7)+p(8)+p(9)+p(10)$

$$= \left(\frac{1}{2}\right)^{10}\left\{\binom{10}{7}+\binom{10}{8}+\binom{10}{9}+\binom{10}{10}\right\} = \frac{120+45+10+1}{1024} = \frac{176}{1024}$$

(B) State and prove central Limit Theorem. (S – 14)

Ans: - **central Limit Theorem.**

If $$X_i = \begin{cases} 1, & with\ probability\ p \\ 0, & with\ probability\ q \end{cases}$$

Then the distribution of the random variable $S_n = X_1 + X_2 + ... + X_n$ where Xi's are independent, is asymptotically normal as $n \to \infty$

Proof: M.G.F. of Xi is given by:

$$M_{X_i}(t) = E(e^{tX_i}) = e^{t \cdot 1} p + e^{t \cdot 0} q = (q + pe^t)$$

M.G.F. of the sum $S_n = X_1 = X_2 + \ldots + X_n$ is given by

$$M_{S_n}(t) = M_{X_1 + X_2 + \ldots + X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \ldots M_{X_n}(t)$$

$$= [M_{X_i}(t)]^n \qquad \text{(since } X_i\text{'s are identically distributed)}$$

$$= (q + pe^t)^n,$$

Which is the M.G.F. of a binomial variate with parameter n and p.

Hence by uniqueness theorem of m.g.f., $S_n \sim B(n,p)$

$\therefore \quad E(S_n) = np = \mu(say), \quad and \quad V(S_n) = nqp = \sigma^2$, (say).

$$Let \quad Z = \frac{E(S_n) - E(S_n)}{\sqrt{VE(S_n)}} = \frac{S_n}{\sigma}$$

$$M_Z(t) = e^{-\mu t/\sigma} M_{S_n}(t/\sigma)$$

$$= e^{-npt/\sqrt{npq}} \left[q + pe^{t/\sqrt{npq}} \right]^n$$

$$= \left[1 + \frac{t^2}{2n} + O(n^{-3/2}) \right]^n$$

Where $O(n^{-3/2})$ represents terms involving $n^{-3/2}$ and higher power of n in the denominator. Proceeding to the limit and $\to \infty$, we get

$$\lim_{n \to \infty} M_Z(t) = \lim_{n \to \infty} \left[1 + \frac{t^2}{2n} + O(n^{-3/2}) \right]^n = \lim_{n \to \infty} \left[1 + \frac{t^2}{2n} \right]^n = e^{t^2/2}$$

Which is the M.G.F. of a standard mormal variate. Hence by the uniqueness theorem of M.G.F.'s

$$Z = \frac{S_n - \mu}{\sigma} \text{ is asymptotically } N(0, 1)$$

Hence Prove.

(C) Explain Geometric distribution (S – 14)

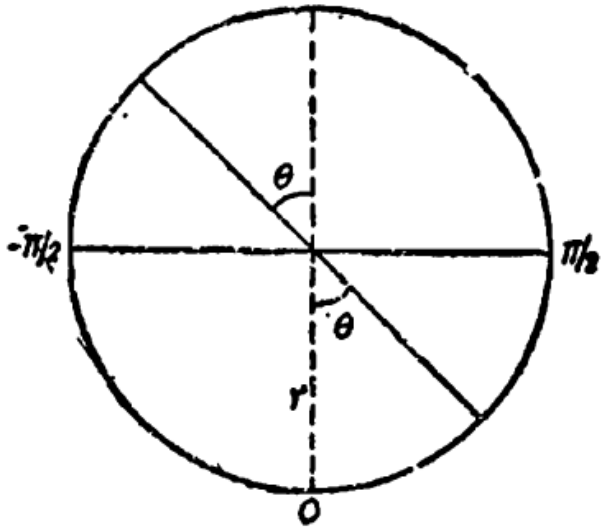
Ans: - Suppose we have a series of independent trials or repetitions and in each trial the probability of success 'p' remains the same. Then the probability that there are x failure preceding the first success is given by q^x p,q = 1-p.

Definition: A random variable X is said to have a geometric distribution if it assumes only non-negative values and its probability mass function is given by:

$$P(X = x) = \begin{cases} q^x p; x = 0,1,2,\dots,; 0 < p \leq 1; q = 1 - p \\ 0 \ otherwise \end{cases}$$

(d) Explain Cauchy distribution with its additive property. (S – 14)

Ans: - Cauchy Distribution : Let us consider a roulette wheel in which the probability of the pointer stopping at any part of the circumference is constant. In other words, the probability that any value of θ lies in the interval [-π/2, π/2] is constant and consequently θ is rectangular variate in the range [-π/2, π/2] with probability differential given by:



$$dP(\theta) = \begin{cases} \left(\frac{1}{\pi}\right) d\theta, -\frac{\pi}{2} \le \theta \le \frac{\pi}{2} \\ 0, \quad otherwise \end{cases}$$

Let us now transform to variable X by the substitute: x=r tan θ => dx = $r\ sec^2\ \theta\ d\theta$

Since, $-\pi/2 \le \theta \le \pi/2$, the range for X is from -∞ to ∞. Thus the probability differential of X becomes:

$$dF(x) = \frac{1}{\pi} \cdot \frac{dx}{r \sec^2 \theta} = \frac{1}{\pi} \cdot \frac{dx}{\left[r \left\{ 1 + (x^2/r^2) \right\} \right]} = \frac{r}{\pi} \cdot \frac{dx}{r^2 + x^2} ; -\infty < x < \infty$$

In a particular if we take r =1 , we get

$$f(x) = \frac{r}{\pi} \cdot \frac{dx}{1 + x^2}, -\infty < x < \infty$$

Definition: A random variable X is said to have a standard Cauchy distribution if it's p.d.f. is given by:

$$f(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty$$

And X is termed as standard Cauchy variate.

More generally, Cauchy distribution with parameter λ and μ has the p.d.f.,

$$g_Y(y) = \frac{\lambda}{\pi[\lambda^2 + (y-\mu^2]}$$

And we write X $\sim$ C(λ, μ)

Q5. (a) State and prove Central limit theorem?

Ans:- **Central limit theorem :-** The central limit theorem in the mathematical theory of probability may be expressed as follows:

"If Xi, (i=1,2,…n) be independent random variable such that $E(X_i) = \mu_i$ *and* $V(X_i) = \sigma_i^2$, then it can be proved that under certain very general condition the random variables Sn = X1 + X2+ ….. +Xn, is asymptotically normal with mean μ and standard deviation σ where

$$\mu = \sum_{i=1}^{n} \mu_i \quad \text{and} \quad \sigma^2 = \sum_{i=1}^{n} \sigma_i^2$$

This theorem was first stated by Laplace and a rigorous proof under fairly general condition was given as a particular case of central limit theorem is De-Moivre's theorem which states as follow:

"If Xi $= \begin{cases} 1, & with\ probability\ P \\ 0, & with\ probability\ q \end{cases}$

Then the distribution of the random variable Sn = X1 + X2 + …….. + Xn where Xi's are independent, is asymptotically normal as n $\to \infty$

Proof: M.G.F. of Xi is given by

$$M_{X_i}(t) = E(e^{tX_i}) = e^{t\cdot 1}p + e^{t\cdot 0}q = (q + pe^t)$$

M.G.F. of the sum $S_n = X_1 = X_2 + \ldots + X_n$ is given by

$$M_{S_n}(t) = M_{X_1+X_2+\ldots+X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \ldots M_{X_n}(t)$$

$$= [M_{X_i}(t)]^n \qquad \text{(since } X_i\text{'s are identically distributed)}$$

$$= (q + pe^t)^n,$$

Which is the M.G.F. of a binomial variate with parameter n and p.

$$\therefore \qquad E(S_n) = np = \mu \text{ (say)}, \text{ and } V(S_n) = npq = \sigma^2, \text{ (say)}.$$

$$\text{Let} \qquad Z = \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - \mu}{\sigma}$$

$$M_Z(t) = e^{-\mu t/\sigma} M_{S_n}(t/\sigma)$$

$$= e^{-npt/\sqrt{npq}} \left[q + pe^{t/\sqrt{npq}} \right]^n$$

$$= \left[1 + \frac{t^2}{2n} + O(n^{-3/2}) \right]^n$$

where $O(n^{-3/2})$ represents terms involving $n^{3/2}$ and higher powers of n in the denominator.

Proceeding to the limits as $n \to \infty$, we get

$$\lim_{n \to \infty} M_Z(t) = \lim_{n \to \infty} \left[1 + \frac{t^2}{2n} + O(n^{-3/2}) \right]^n = \lim_{n \to \infty} \left[1 + \frac{t^2}{2n} \right]^n = e^{t^2/2}$$

which is the M.G.F. of a standard normal variate. Hence by the uniqueness theorem of M.G.F.'s

$$Z = \frac{S_n - \mu}{\sigma} \text{ is asymptotically } N(0, 1).$$

Hence $S_n = X_1 + X_2 + \ldots + X_n$ is asymptotically $N(\mu, \sigma^2)$ as $n \to \infty$.

(b) Explain gamma distribution with its cumulant generating function.

Ans: - Gamma Distribution:- The continuous random variable X which is distributed according to the probability law:

$$f(x) = \begin{cases} \frac{e^{-x} x^{\lambda-1}}{\Gamma(\lambda)}; \ \lambda > 0, 0 < x < \infty \\ 0, otherwise \end{cases}$$

Is known as a Gamma variate with parameter λ and referred to as a $\gamma(\lambda)$ variate and its distribution is called Gamma distribution.

Cumulant Generating function of Gamma Distribution:- The cumulant generating function Kx(t) is given by

$$K_X(t) = \log M_X(t) = \log (1 - t)^{-\lambda} = -\lambda \log (1 - t); \ |t| < 1$$

$$= \lambda \left[t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \cdots \right]$$

$\therefore$ Mean $= \kappa_1 =$ Coefficient of t in $K_X(t) = \lambda$

$\mu_2 = \kappa_2 =$ Coefficient of $\frac{t^2}{2!}$ in $K_X(t) = \lambda$

$\kappa_3 =$ Coefficient of $\frac{t^3}{3!}$ in $K_X(t) = 2\lambda$

$\kappa_4 =$ Coefficient of $\frac{t^4}{4!}$ in $K_X(t) = 6\lambda$

$\therefore$ $\mu_4 = \kappa_4 + 3\kappa_2^2 = 6\lambda + 3\lambda^2$

Hence $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{4\lambda^2}{\lambda^3} = \frac{4}{\lambda}$ and $\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{6}{\lambda}$

(c) Find the mode and median of Normal Distribution

Ans:- Mode of Normal Distribution:- Mode is the value of x for which f(x) is maximum, i.e., mode is the solution of

f'(x) = 0 and f''(x) <0

for normal distribution with mean μ and standard deviation σ,

$$\log f(x) = c - \frac{1}{2\sigma^2}(x-\mu)^2,$$

Where c = $\log(1/\sqrt{2\pi\sigma})$, is constant.

Differentiating w.r.t. x we get

$$\frac{1}{f(x)} \cdot f'(x) = -\frac{1}{\sigma^2}(x-\mu) \Rightarrow f'(x) = -\frac{1}{\sigma^2}(x-\mu) f(x)$$

And

$$f''(x) = -\frac{1}{\sigma^2}\left[\, 1 \cdot f(x) + (x-\mu) f'(x) \,\right] = -\frac{f(x)}{\sigma^2}\left[\, 1 - \frac{(x-\mu)^2}{\sigma^2} \,\right]$$

Now f'(x) = 0 => x - μ = 0 i.e., x = μ

At the point x = μ, we have from equation

$$f''(x) = -\frac{1}{\sigma^2}[f(x)]_{x=\mu} = -\frac{1}{\sigma^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} < 0$$

Hence x = μ, is the mode of the normal distribution.

Median of Normal Distribution:- If M is the median of the normal distribution, we have

$$\int_{-\infty}^{M} f(x)\,dx = \frac{1}{2} \Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{M} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = \frac{1}{2}$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp\left\{-(x-\mu)^2/(2\sigma^2)\right\} dx = \frac{1}{2}$$

But

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\mu}\exp\left\{-(x-\mu)^2/2\sigma^2\right\}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0}\exp\left(-z^2/2\right)dz=\frac{1}{2}$$

From equation we get

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\mu}\exp\left\{-(x-\mu)^2/2\sigma^2\right\}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0}\exp\left(-z^2/2\right)dz=\frac{1}{2}$$

$$\Rightarrow \qquad \frac{1}{\sigma\sqrt{2\pi}}\int_{\mu}^{M}\exp\left\{-(x-\mu)^2/2\sigma^2\right\}dx=0 \quad \Rightarrow \quad \mu=M$$

Hence for the normal distribution Median.