PROBABILITY AND STATISTICS

For Enginerring

4th Semester

CTEVT

Website:-www.arjun00.com.np

Website :- https://www.arjun00.com.np

Unit-1

Introduction to statics

- **Origin of statistics**

The term "statistics" appears to have been derived from the Latin word "Status" or Italian word "Statista" or German word "Statistik" all of which mean the political state . In those day, "statistics" was used only in collecting the information relating to the population of the state, military strength, income etc for framing the military and the fiscal policies.

- So , at that time **"statistics"** was considered only as the science of state craft. However with passage of time ,the science of statistics has been applied very widly.So , in the modern time , the scope of statistics has considerably enlarged. It is used not only to the state in administration but it is used in Economics , Science , Business , research, Bank etc .

Website :- https://www.arjun00.com.np

Meaning of Statistics

- The word "Statistics" issued in singular as well as plural sence . In the plural sense , it means the quatitative information or numerical facts collected systematically. As for example: statistics of population, national incomes , unemployment , poverty, exports ,imports etc . In the singular sence , it means the various methods and techniques adopted for the collection, presentation, analysis and the interpretation of the figures .

Definition of Statistics

Definition according to singular sense:

- According to the Croxton and cowden have given a very simple and concise definition of statistics . In their view "statistics may be defined as the collection , presentation , analysis and interpretation of numerical data ".
- According to the Prof. Ya-Lun-Chu : "statistics is a method of decision making in the face of uncertainity on the basis of numerical data and calculated risks "

Website :- https://www.arjun00.com.np

Definition according to plural sense

- Yule and Kendall define that " By statistics, we mean quantitative data affected to a marked extend by a multiplicity of cause "
- Horace Secrist define that " By statistics, we mean aggregate of facts affected to a marked extend by multiplicity of cause, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre –determined purpose and placed in relation to each other .

Important and scope of statistics

- It is useful in the administration of various states .
- It is useful in the businessman .
- It is useful in the economists
- It is useful in research workers .
- It is useful in planning .
- It is useful in the bankers .
- It is useful in the insurances companies .
- It is useful in the supervision .
- It is useful in physics , biological and other science

Website :- https://www.arjun00.com.np

Function of Statistics

- It presents facts in a definite form .
- It simplifies complexity .
- It facilitates comparison .
- It helps in formulating and testing hypothesis .
- It helps in prediction (forecasting) .
- It helps in formulation of suitable policies .
- It helps other sciences in research and conducting test .

Limitation of Statistics

- Statistics does not deal with individuals .
- Statistics does not study qualitative phenomena.
- Statistics laws are not exact .
- Statistics is only a mean .
- Statistics is liable to be misused .

Home Assignment

- ***Q:-*** Define the term statistics and write the function of statistics. (CTEVT 2073)

Website :- https://www.arjun00.com.np

Unit – 2

Basic Concept of Statistical Studies

- **Introduction :** This unit gives the basic idea of data, its types , idea of collecting data from different sources. To differentiate between population and sample , parameter and statistics . Also its give the basic concept of variables and its types which are very important for the study of statistics .

Data

- From the very definition of statistics , it is an aggregate of facts that can numerically be expressed . These types of facts are known as data . For drawing any statistical inference we need to collect and complete them .

Collection of Data

For any statistical investigation , the first attempt is to collect the data . Before collecting the data for a given statistical enquiry , it is essential to examine carefully the following points :-

- Objectives and scope of the inquiry .
- Sources of information .
- Method of data collection .
- Units of data collection .
- Degree of accuracy aimed a in the final results .

Website :- https://www.arjun00.com.np

Types of Data

Depending on the sources, statistical data are classified under two categories and they are primary and secondary data .

- ***Primary data :-*** The data that are originally collected by an investigator for the first time for the purpose of statistical enquiry are taken as primary data . For example: data obtained in a population census by the ministry of population and environment in every ten years is primary data .

Secondary data

The data that are not originally collected but rather obtained from some published or unpublished sources are known as secondary data . For example: the reports and publications made by central bureau of statistics are primary for that organization but secondary data for those who use it .

Website :- https://www.arjun00.com.np

Sources of Data

Sources of Primary Data

- Direct personal interviews : Under this method of collecting data , there is a face to face contact with the person from whom the information is to be obtained .
- Indirect and interviews : Under this method , the investigator contact third parties called witnes capable of supplying the nessessary information. This method is generally adopted in case when the informants hesitate to give information directly .
- Informatiom from correspondents:

  Under this method , the investigator appoints local agents or correspondents in different places to collect information . These agents collect the information and supply it to the central office , where this information is further processed.
- Mailed questionnaire method :

  Under this method , a list of questions pertaining to the enquiry is prepared and is send by post to the various persons from whom the information is to be collected .

Website :- https://www.arjun00.com.np

Sources of Secondary Data

The sources of secondary data can broadly be classified under two heads and they are published sources and unpublished sources.

Published sources : Published data are found from various sources such as :

- Official publications of different ministries and departments of the government .
- Reports and publications of organizations like Nepal chamber of Commerce , Nepal Rastra bank etc.
- Reports of economists , research scholars, universities and so on .
- Official publication of foreign government or international bodies .

Unpublished sources :

- All statistical material is not always published . There are various sources of unpublished data such as records maintained by various government and private offices , studies made by research institutions , scholars etc .

Website :- https://www.arjun00.com.np

Population and Sample

- The whole mass of observation (data) collecting from different sources is called the population(or universe) . That is the total number of data collected are called population .
- The randomly selected items (pieces) from the whole observation are called the sample pieces or simply it is called sample . And the process of selecting some of the sample pieces from the whole mass of observation is called sampling .

Parameter and statistic

- The statistical measures like mean, median, standard devition, moments etc . which are computed from population (or universe) are called parameters . Parameter in statistics is any numerical quantity that characterizes a given population or some aspect of it .
- Again the statistical meaasures computed from the sample data are called the statistics .

Variables

Any quantity to show variation from one individual to the other in the same population is known as variable . An individual observation of any variable is known as variate . A variable may be either discrete or continuous .

- ***Discrete Variable*** : A discrete varible is that which can not taken all value and there is a gap between one value and the other . The discrete variable may take any integer value from 0 to ∞ .

Continuous Variables :

- These variables which can take all the possible values in a given specified range are called continuous variable .

Home Assignment

- Q :- Distinguish between primary and secondary data . What are the sources of secondary data ? (CTEVT 2072/2074/2075)

Unit – 3

Organizing a Raw Data

Meaning of classification:

Classification of data means the process of arranging the data into different groups of classes according to their similarities . Thus classification is the first step in tabulation because the items with similarities must be brought together before the data are presented in the form of table.

Objectives of classification

- It presents the facts into a simpler form and pinpoints the significant features of data at a glance .
- It condenses the huge mass of data by eliminating irreverent details such that points of similarities and dissimilarities can easily be brought out .
- It facilitates comparison .
- It helps to enable a statistical treatment of information collected.
- It prepares the basis of tabulation .

Website :- https://www.arjun00.com.np

Types of Classification

- The collected data can be classified according to the basis of following four ways :
- **Geographical:-** Area wise (regional,cities, districts etc.)
- **Chronological:-** On the basis of time.
- **Qualitative:-** On the basis of some character .
- **Quantitative:-** In terms of magnitude .

Frequency and Frequency Distribuion

- ***Frequency* :-** The number of times which a variate value occurs is known as frequency . The variate value 5 repeats 7 times means the frequency of the variate value 5 is 7 .
- ***Frequency Distribution:-*** The arrangement of the value of variable and their corresponding frequencies is called frequency distribution . The table in which the value of variable and their corresponding frequencies are arranged is called frequency distribution table.

Types of Frequency Distribution

- There are two types of frequency distribution:
- **1. Univariate :** univariate frequency distribution may be of the following three types :
- Individual series
- discrete series
- Continuous series
- **2. Bivariate:** If the frequency distribution consists two variable then it is called bivariate frequency distribution

1. Individual frequency distribution

In this frequency distribution, the individual value of variable is listed. The distribution is used when the number of observation is small and all observation are distinct . For example the temperature of a certain city in the first week of august is given as follows :

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat
0C	34	33	35	37	31	33	32

2. Discrete series

- The series formed from a discrete variable is known as discrete series . In a discrete series , the data are presented in a way that exact measurements of units are clearly indicated. While writing the frequency we use the tally marks . For example, we make following frequency distribution from 25 women according to their number of child

3,2,3,4,2,2,3,4,5,1,1,2,3,4,3,2,2,2,1,4,3,2,2,3,1

No. of children	No of women	
X	Tally Marks	Frequency (f)
1	I I I I	4
2	IIII IIII	9
3	IIII II	7
4	IIII	4
5	I	1
	Total	=25

Website :- https://www.arjun00.com.np

Continuous Frequency Distribution

When set of observations contains large number of distinct values then the distribution of data in discrete series will be much complicated . In this case , we form certain group of variate value called classes . Then we count the number of variate value lying inside the class as the frequency of that class .

For example the mark of 28 students in a test is given :

95, 72 , 05 , 09 , 11 , 18 , 25 , 29 , 21 , 33 , 55, 67 , 73 , 27 , 44 , 49 , 93 , 82 , 86 , 33 , 51 , 63 , 45 , 32 , 33 , 39 , 42 , 89 .

Class	Frequency
0-10	2
10-20	2
20-30	4
30-40	5
40-50	4
50-60	2
60-70	2
70-80	2
80-90	3
90-100	2

Classification according to the class-interval

- **Class :-** Different groups into which the data can be classified are known as classes or class-interval.
- **Class Limit :-** The two values specifying the class are called the class limits. The smaller value is said to be the lower limit and the greater value is called the upper limit.
- **Class size:-** The difference between the upper and lower limits of a class is called magnitude of the class size .
- **Class Frequency:-** The number of observations falling within a particular class is known as the class frequency .
- **Mid-value of the class:-** It is the value lying in the halfway between the two extreme limits of a class . It is also known as mid-point or class mark . It is determined by the following method:

  The mid-value of the class interval 30-40 is

  (30 + 40)/2 =35 .

Method of forming class-intervals

- There are two ways of forming class-interval :
- **(i) Exclusive method :-** The formation of class-interval by this method is that the upper limit of one class is the lower limit of the next class so as to make continuous without any gap .

Rainfall(in mm.):	10-20	20-30	30-40	40-50	50-60
No. of days:	12	7	10	5	2

(ii) Inclusive method

The formation of the class-interval by this method is that both lower and upper are included in a particular class .

Rainfall(in mm.):	10-19	20-29	30-39	40-49	50-59
No. of days:	12	7	10	5	2

Website :- https://www.arjun00.com.np

Correction Factor

- When the class intervals are formed by "inclusive method" there is a gap between the upper limit of any class and the lower limit of succeeding class . To change the classes of "inclusive type" into that of "exclusive type" we adjust and get the real limit of the classes by using the correction factor . The correction factor denoted by C_f

$$C_f = \frac{Lower\ \lim of\ 2ndclass - Upper\ \lim of\ 1stclass}{2}$$

Then the real lower limit =Lower limit - C_f

And the real upper limit =Upper limit + C_f

Example: Change the following frequency distribution of inclusive type into exclusive type

Rainfall in mm.	No. of days
10-19	12
20-29	7
30-39	10
40-49	4
50-59	2

Website :- https://www.arjun00.com.np

Solution:

Correction factor = $C_f = \frac{Lower\lim of\ 2ndclass}{}$

= (20 – 19) / 2 = 0.5

Rainfall in mm.	No. of days
9.5-19.5	12
19.5-29.5	7
29.5-39.5	10
39.5-49.5	4
49.5-59.5	2

Cumulative Frequency Distribution

The frequency distribution in which frequencies are added is called cumulative frequency distribution. They are mainly two types of cumulative frequency distribution .

- Less than cumulative frequency distribution
- More than cumulative frequency distribution

Website :- https://www.arjun00.com.np

(a) Less than cumulative frequency distribution

To make less than cumulative frequency distribution, we arrange the classes in the ascending order of magnitude and we add the frequencies from the top.

(b) More than cumulative frequency distribution

In this case , the first step is arranging the data in their ascending order of magnitude and we add the frequency from bottom .

Example : From the following frequency distribution ,make less than and more than cumulative frequency distribbution

Marks	No. of students
0- 10	3
10-20	7
20-30	11
30-40	5
40-50	8

Solution : .....

The less than cumulative frequency distribution table

Marks	No, of students
Less than 10	3
Less than 20	3 + 7 = 10
Less than 30	10 + 11 = 21
Less than 40	21 + 5 = 26
Less than 50	26 + 8 = 34

More than cumulative frequency distribution table

Marks	No, of students
More than 0	34
More than10	34 – 3 = 31
More than 20	31 – 7 = 24
More than 30	24 – 11 = 13
More than 40	13 – 5 = 8

Website :- https://www.arjun00.com.np

Tabulation of Data

Tabulation is the process of orderly arrangement of data into series of rows and columns . Rows are horizontal arrangements and columns , the vertical arrangements. The purpose of the tabulation is to simplify the presentation of data and to facilitate comparison between related information. It is the final step of collection of data .

Different parts of a table

- The different parts of table depend upon the nature of the data and the purpose of investigation . Generally, the main parts of a table are mentioned below :

Table number
2) Title of the table
3) Caption
4) Stubs
5) Body of the table
5) Head note
6) Footnote
7) Source

Website :- https://www.arjun00.com.np

Format of a table

Table No.....

Title.....

Head note:- ....

Captions (title)			
Stubs	Sub heads	Sub heads	Total
Total	Body		

Foot note:-

Source note:-

Types of Table

Generally there are two types of table which are classified according to the following basis:

- On the basis of construction
- On the basis of purpose

1) On the basis of construction:

a) Simple b) Complex

2) On the basis of purpose:

a) General purpose b) Special purpose

Website :- https://www.arjun00.com.np

Example:

In a simple study about the smoking habits in two towns, following data were derived:

	Town A	Town B
Male in total population:	52%	54%
Smokers:	26%	28%
Male Smokers:	18%	20%

Tabulate the above data .

Solution:

Smoking habits of two town A and town B

Attribute	Town A			Town B		
	M	F	T	M	F	T
Smokers	18	8	26	20	8	28
Non-smokers	34	40	74	34	38	72
Total	52	48	100	54	46	100

M = Males F = Females T = Total = 100

Website :- https://www.arjun00.com.np

Example:

In a sample study about coffee drinking habits in two towns , the following information was received.

- Town A : Female were 40% , total coffee drinkers were 45% , and male non coffee drinkers were 15% .
- Town B: Male were 55% , male non-coffee drinkers were 30% , and female coffee drinkers were 15% .

Solution:

Coffee drinking habits of town A and town B

Attribute	Town A			Town B		
	M	F	T	M	F	T
Coffee drinkers	45	0	45	25	15	40
Non- Coffee drinkers	15	40	55	30	30	60
Total	60	40	100	55	45	100

M = Males F = Females T = Total = 100

Website :- https://www.arjun00.com.np

Exercise 3

1. What do you mean by "classification of data" ? What are the objectives of classification in statistics ?

2. Define classification. Explain the various ways of classification adopted in statistics ?

3. Discuss the general rules of classification of data.

4. What is a variable ? What is the difference between discrete and continuous variables ?

5. Define tabulation. Distinquish between classification and tabulation.

6. In an enquire of 30 families, the number of children recorded is as follows:

   2, 1, 0, 3, 2, 2, 4, 1, 2, 3, 4, 2, 5, 0, 2,
   3, 2, 4, 0, 3, 2, 3, 2, 4, 1, 2, 3, 4, 5, 1,

   Represent the above data into discrete frequency distribution.

7. The following is the set of marks out of 50 obtained by 40 students:

   22 27 9 22 31 17 13 15 18 3 1 21 27 28
   15 13 34 0 34 29 26 16 25 33 36 10 24 22
   26 19 14 36 18 25 21 33 35 25 18 18

   Form a frequency distribution by taking a class interval of 5.

8. Prepare a frequency distribution taking a class interval of 10 from the following figures relating to the wages in rupees of 50 workers:

   70 55 51 42 57 45 60 47 63 53 33 65 39
   82 55 64 58 61 65 42 50 52 54 45 45 25
   36 59 63 39 65 45 49 54 64 75 42 41 52
   35 30 35 15 48 26 20 40 55 46 18

9. Form a frequency distribution from the following data by "Inclusive method" taking 3 as the magnitude of class intervals:

   10 17 15 22 11 16 19 24 29 18 25 26 32 14
   17 20 23 27 30 12 15 18 24 36 18 15 21 28
   33 38 34 13 10 16 20 22 29 16 23 31

10. Construct an ordinary frequency table from the following data:

(i)

Income (in Rs)	No. of persons	Income (in Rs)	No. of persons
Above 0	100	Above 500	23
Above 100	96	Above 600	15
Above 200	76	Above 700	3
Above 300	49		
Above 400	35		

(ii)

Daily wages (in Rs)	No. of workers	Daily wages (in Rs)	No. of workers
Below 20	10	Below 60	89
Below 30	34	Below 70	97
Below 40	61	Below 80	100
Below 50	78		

11. In a sample study about coffee drinking habits in two towns, the following information was received.

    Town A: Female were 40%, total coffee drinkers were 45% and male non-coffee drinkers were 20%. 15%

    Town B: Males were 55% male non-coffee drinkers were 30% and female coffee drinkers were 15%.

    Present the above information in a suitable form.

Website :- https://www.arjun00.com.np

Answers

6.

No of children:	0	1	2	3	4	5
No of families:	3	4	10	6	5	2

7.

Marks:	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
No of students:	2	1	5	9	6	8	5	4

8.

Wages (in Rs.):	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No of workers:	2	3	7	13	13	9	2	1

9.

Class:	10-13	14-17	18-21	22-25	26-29	30-33	34-37	38-41
Frequency:	5	8	8	7	5	4	2	1

10. (i)

Income	No. of persons
0-100	4
100-200	20
200-300	27
300-400	14
400-500	12
500-600	8
600-700	12
700-800	3

(ii)

Daily wages	No. of workers
10-20	10
20-30	24
30-40	27
40-50	17
50-60	11
60-70	8
70-80	3

11. COFFEE DRINKING HABITS OF TOWN "A" AND "B"

(in percentage)

Attribute	Town A			Town B		
	Male	Fem.	Total	Male	Fem.	Total
Coffee drinkers:	45	-	45	25	15	40
Non Coffee drinkers:	15	40	55	30	30	60
Total:	60	40	100	55	45	100

Unit:4

Pictorial Representation of Data Set

Introduction:

After collecting the data , we classify and tabulate them . But the representation of data in table only is not always sufficient to understand and analyze the data .To avoid these difficulties we present the data in graphs and diagrams . Graphs and diagrams are the visual form of presentation of data where comparisons can be made between two or more phenomenon .In graphic and diagrammatic representation , we use geometric figure like points , lines, rectangles, bars and circles instead of numerical value

Website :- https://www.arjun00.com.np

Importance of Diagrammetic and Graphical representation

- Diagrams and graphs gives a birds eye's view of a set of numerical data. They can present than the data into simple and intelligible form.
- Diagrams are generally more attractive and impressive the numerical data . They give delight to eye and leave an ever lasting impression on the mind .
- Diagrams help in deriving the required information in less time and without any mental strain .
- They facilitate comparison of two or more set of data at a time .

Difference between diagrams and graphs

- Diagrams are constructed on plane paper whereas the graphs are on graph the paper .
- Diagrams are used only for the comparison but graphs help in studying the mathematical relationship between two variables .
- In diagrams, the numerical data are presented by bars ,rectangles ,circles ,cubes ,etc whereas in graphs the data are presented in terms of pints and lines .

Website :- https://www.arjun00.com.np

- Diagrams may be one ,two or three dimensional but graphs , will generally be of two dimensional.
- Construction of diagrams are not so easy but the construction of graphs is easier in comparison to diagrams .
- Presentation of frequency distribution in diagrams is not used but the presentation of frequency distribution and time series in graphs is more appropriate .

Diagrammatic Representation of Data

In diagrammatic analysis there are two types of diagrams are commonly used (1) Bar diagram (2) Pie-Chart (Angular Diagram).

(1) Bar Diagram :- Bar diagram are also known as one dimensional diagram in which height of the bars represents the magnitude of the data .Bar diagram consists of set of rectangles of equal width and equal spacing between rectangles.

Types of bar diagram

- Simple bar diagram.
- Sub divided or component bar diagram.
- Percentage bar diagram .
- Multiple bar diagram .

Simple bar diagram

Simple bar diagram is a diagram for only one variable . It is the most simple and frequently used for the comparative study of two or more values of a single variable .

Example :1

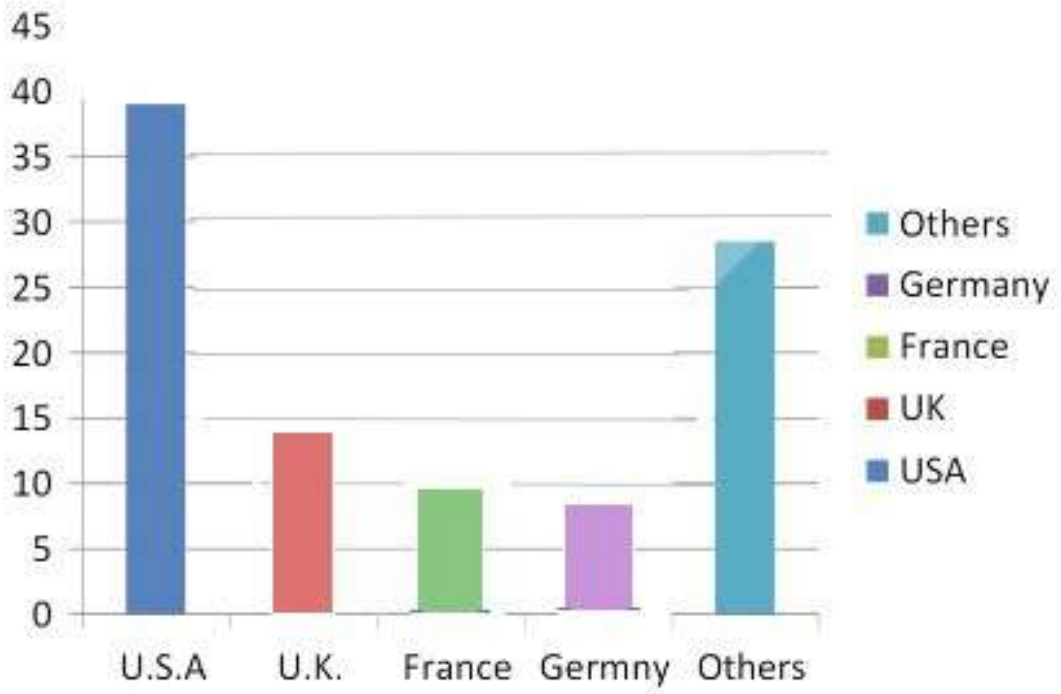
- The following information gives the percentage of tourists by major nationality in the year 1968 for the kingdom of Nepal:

Nationality	Percentage of tourists
U.S.A	39.4
U.K.	13.2
France	10.0
Germany	8.9
Others	28.5

Represent the above information by simple bar diagram

Website :- https://www.arjun00.com.np

Solution:



Sub-divided or component bar diagram

Sub-divided bar diagram is a diagrams for two or more components of a total.

Example :2 The following table gives the lengths of roads in km. and its type in different zones

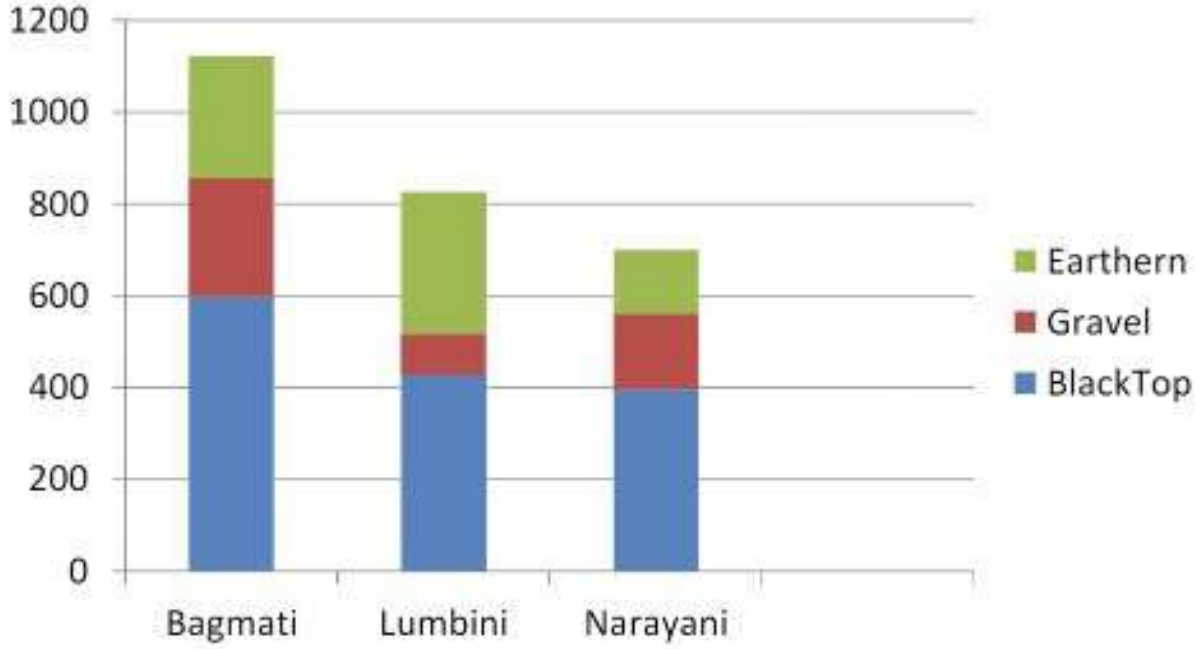
Zone	black-top	Gravel	Earthern	Total length
Bagmati	598	258	266	1122
Lumbini	428	90	308	826
Narayani	396	165	139	700

Represent the above data by sub-divided bar diagram.

Website :- https://www.arjun00.com.np

Solution:

Sub divided bar diagram of lengths of road



Percentage bar diagram

A sub- divided bar diagram presented in a percentage basis is known as percentage bar diagram.

Example:3 percentage break up of the cost of constructing a house in Kathmandu is given below :

Labour = 25% Bricks = 15% Cement = 20%

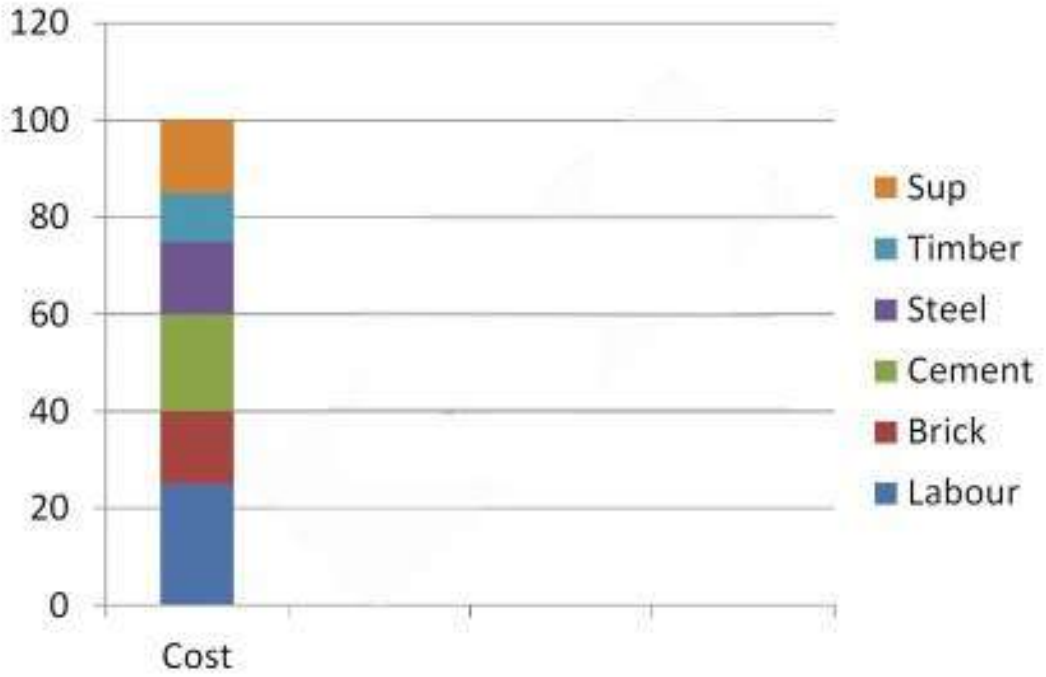
Steel = 15% Timber = 10% Supervision = 15%

Website :- https://www.arjun00.com.np

Solution:

Item	% Expenditure	Cum % Exp
Labour	25	25
Bricks	15	40
Cement	20	60
Steel	15	75
Timber	10	85
Supervision	15	100

Division of cost in construction a house



Website :- https://www.arjun00.com.np

Multiple bar diagram

- It is also a one dimensional bar diagram. When two or more variables are to be computed as the same time, we draw multiple bar diagram. In a multiple bar diagram , two or more sets of inter-related data are represented. In a multiple bar diagram , a set of adjacent bars are drawn .

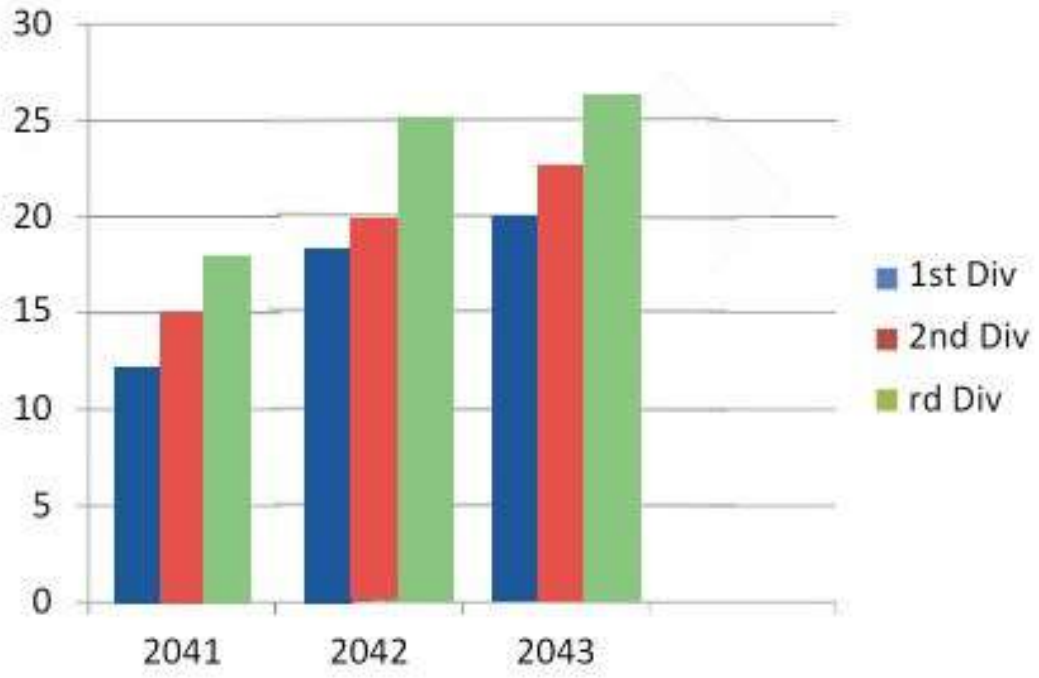
Example:4

- The following table shows the results of S.L.C. examination of a certain school in 3 years. Represent by multiple bar diagram.

Year	1st. Div.	2nd Div.	3rd Div.
2041	12	15	18
2042	17	20	25
2043	20	22	26

Website :- https://www.arjun00.com.np

Solution:



Pie-Chart or Angular Diagram

- It is a type of two dimensional diagram in which area represent the total value. In pie –chart ,we construct a circle whose area represent the total value of variable . The circle is divided in to different sectors by the radial lines such that area of each sector represent the value of components. In pie-chart , area of the sector denotes the frequency of the variable . The size of angle is obtained by multiplying the class percentage by 3.6 or use the formula

$$Size\ of\ angle = \frac{Class\ frequency}{Totalo\ bservation} \times 360$$

Website :- https://www.arjun00.com.np

Example:5

- Draw a pie-chart for the following percentage component of a family expenditure:

Food	= 60%
Rent	= 15%
Clothing	= 10%
Education	= 8%
Health	= 5%
Other item	= 2% .

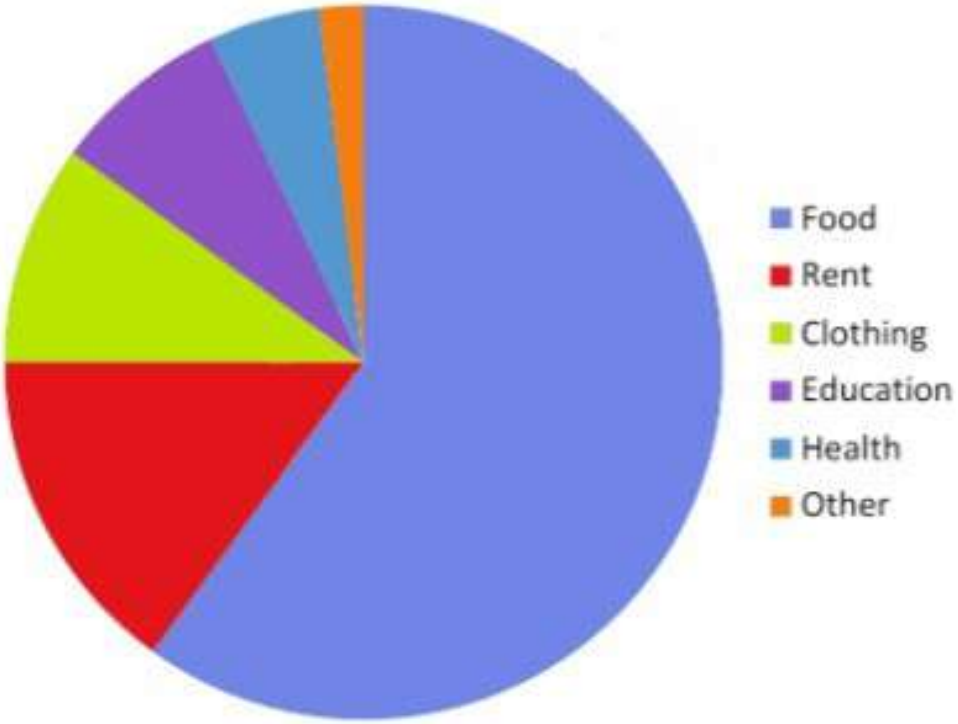
Solution:

- Taking 100 % = 360^0
  1% = 3.6^0

Item	% Expenditure	Angle at centre
Food	= 60%	60 x 3.6 = 216
Rent	= 15%	15 x 3.6 = 54
Clothing	= 10%	10 x 3.6 = 36
Education	= 8%	8 x 3.6 = 28.8
Health	= 5%	5 x 3.6 = 18
Other item	= 2%	2 x 3.6 = 7.2
	100	360

Website :- https://www.arjun00.com.np

Pie-chart of the composition expenditure



Example:6

- Compare the monthly expenditure of two families given below with the help of angular diagrams:

Items of exp.	Family A (Income Rs. 500)	Family B (Income Rs.720)
Food	200	240
Clothing	80	160
House Rent	100	120
Education	30	80
Fuel and lighting	40	40
Miscellaneous	50	80

Website :- https://www.arjun00.com.np

Solution:

For Family A

Taking Rs. 500 = 360°

Rs.1 = $\frac{360^\circ}{500} = 0.72^\circ$

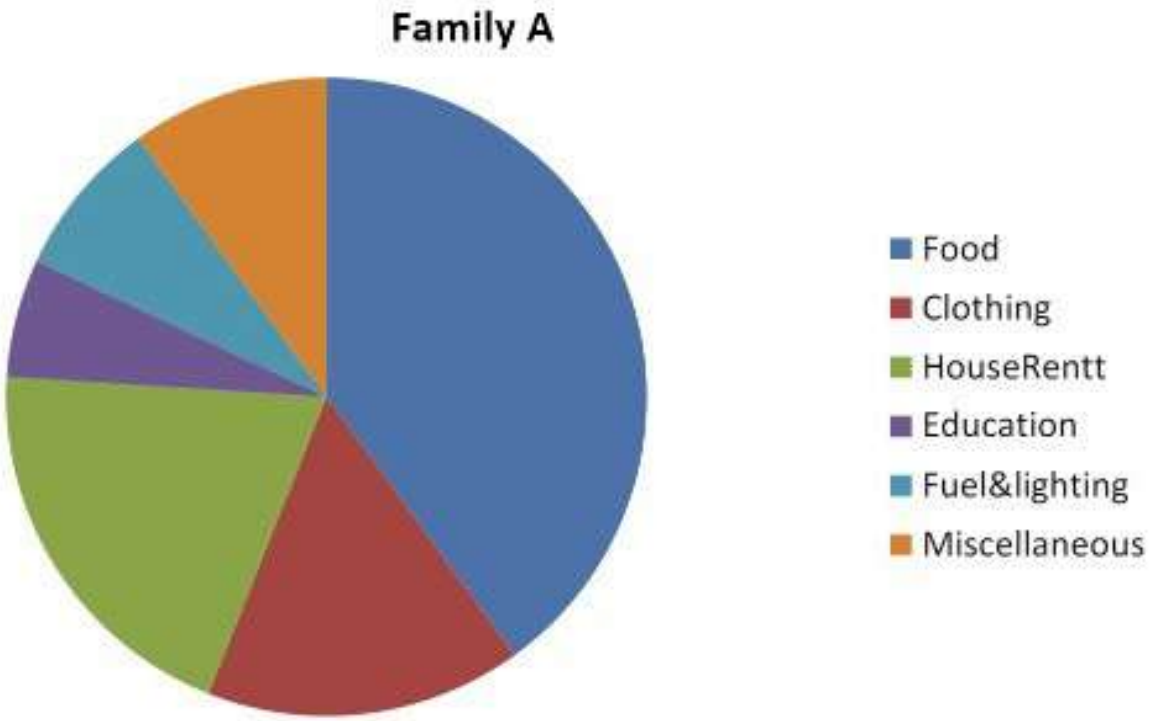
For Family B

Taking Rs.720 = 360°

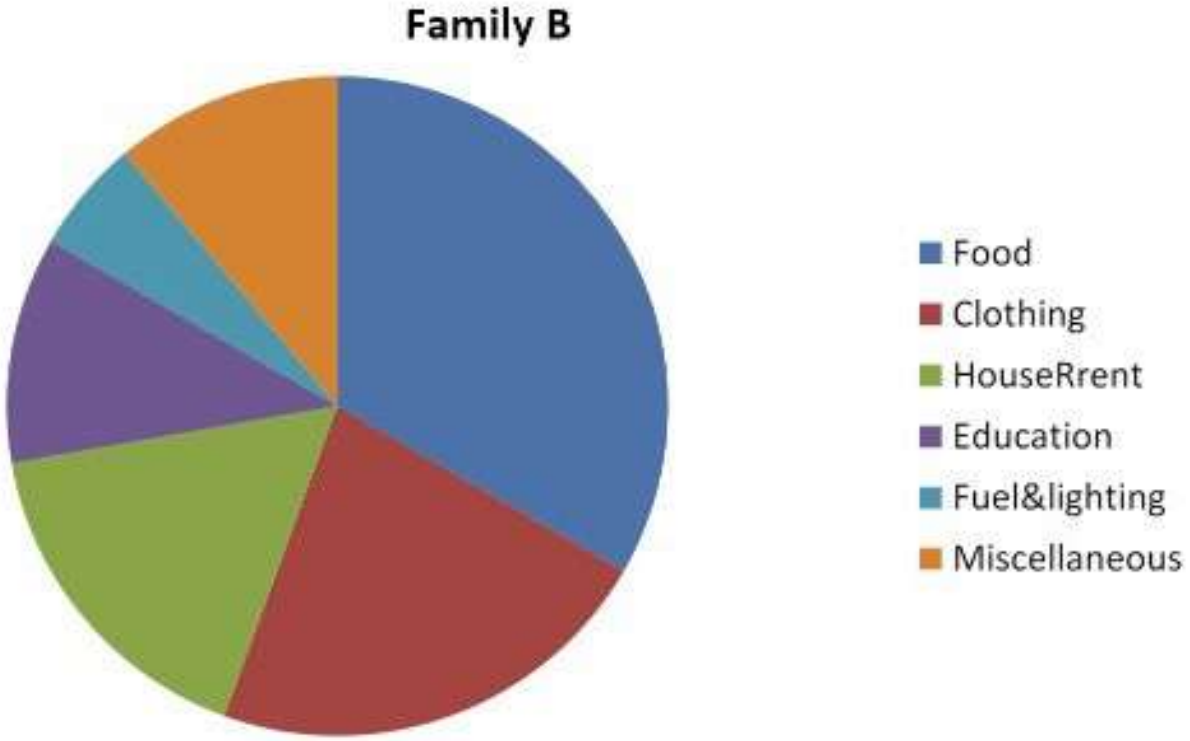
Rs. 1 = $\frac{360^\circ}{720} = 0.5^\circ$

Items	Family A		Family B	
	Exp.	Angle at centre	Exp.	Angle at centre
Food	200	200 x 0.72 = 144	240	240 x 0.5 = 120
Clothing	80	80 x 0.72 = 57.6	160	160 x 0.5 = 80
House Rent	100	100 x 0.72 = 72	120	120 x 0.5 = 60
Education	30	30 x 0.72 = 21.6	80	80 x 0.5 = 40
Fuel and lighting	40	40 x 0.72 = 28.8	40	40 x 0.5 = 20
Miscellaneous	50	50 x 0.72 = 36	80	80 x 0.5 = 40
Total	500	360	720	360

Pie-Chart Of Monthly Exp Of Two Family



Website :- https://www.arjun00.com.np



Graphical Representation Of Data

Website :- https://www.arjun00.com.np

Graphical Representation of Data

- In graphical method we use graph paper where we draw two straight lines , one horizontal and other vertical line. The horizontal line is called abscissa (X-axis) and the vertical line is called ordinate (Y-axis) , the meeting point is called origin .
- To represent the data in graph we first change them into grouped data i.e. in the frequency distribution table. The grouped data so obtained can be represented graphically in any one of the following ways:-
  (a) Histogram
  (b) frequency polygon
  (c) frequency curve
  (d) Cumulative frequency curve or Ogive

(a) Histogram

- Presentation of frequency distribution in the graphical form known as "Histogram" . It consists of a set of adjecent vertical rectangles on x-axis with bases equal width of the corresponding class interval and heights proportional to the corresponding frequency of each class.

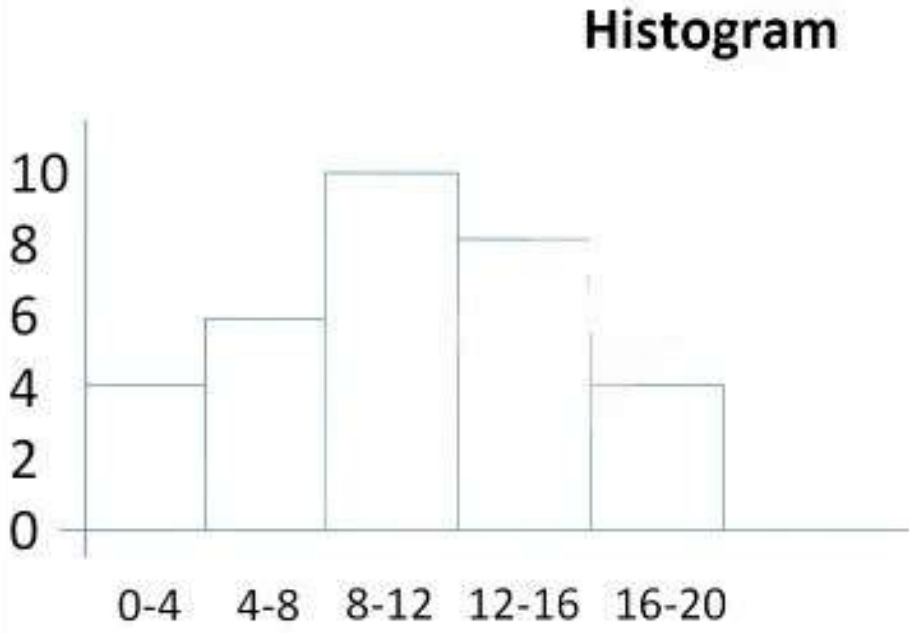
Example:7 Draw histogram for the following data:

Marks:	0-4	4-8	8-12	12-16	16-20
No. of students:	4	6	10	8	4

Website :- https://www.arjun00.com.np

Solution:

- Here class sizes are equal, so height are proportional to the corresponding frequencies.



(b) Frequency Polygon

- Frequency polygon is another method of graphical representation of frequency distribution. When the distribution is discrete, the frequency polygon obtained by plotting the points with values of the variable as the x-coordinate and the corresponding frequency as the y-coordinate and joining these points by straight lines
- If the frequency distribution is continuous then the frequency polygon can be drawn from the following two ways:-

Website :- https://www.arjun00.com.np

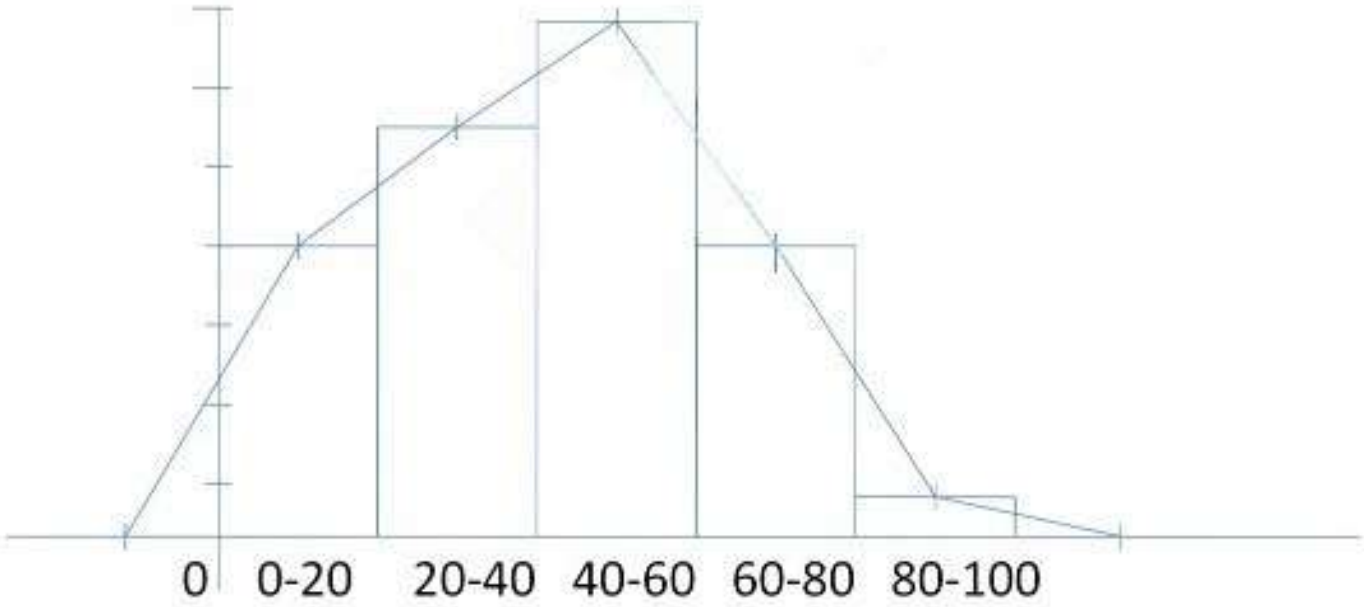
- (a) Using histogram:- First draw the histogram of the given frequency distribution . Find the middle points of the top of the adjecent rectangles. Join these middle points by a straight lines. The figure thus obtained is known as frequency polygon.
- (b) Without using histogram:- First find the middle points of every class-interval. Consider the middle points as x-coordinate and the points are plotted as in discrete frequency distribution.

Example:8

- Draw the histogram and then the frequency polygon from the following data :

Weekly wages:	0-20	20-40	40-60	60-80	80-100
No. of student:	40	51	64	38	7

Solution:- **Histogram and Frequency Polygon**



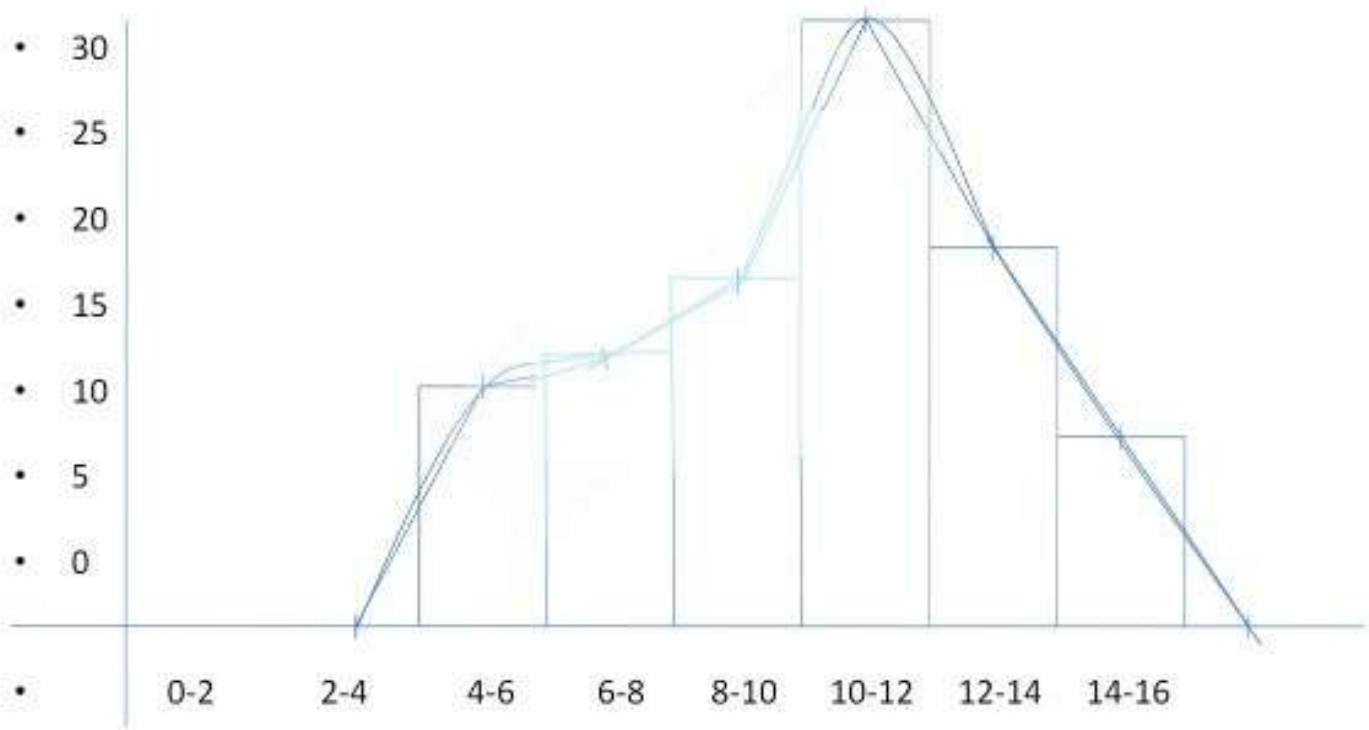
Website :- https://www.arjun00.com.np

(c) Frequency Curve

- A smooth free hand curve drawn through the vertices of a frequency polygon is known as the frequency curve .
- **Example:9** Draw a histogram, frequency polygon and the frequency curve representing the following figure:

Lengths of leaves (cms)	4-6	6-8	8-10	10-12	12-14	14-16
No. of leaves	10	12	16	30	18	8

Histogram, Frequency Polygon and Frequency curve



Website :- https://www.arjun00.com.np

(d) Cumulative Frequency Curve Or Ogive

- Cumulative frequency curve which is also known as "Ogive" is the graphical representation of the cumulative frequency distribution of a continuous variables. In drawing ogive , points are plotted with cumulative frequency (c.f.) along the y-axis and the corresponding class boundaries along the x-axis and joining them freely. There are two types ogives:

  (i) Less than ogive (ii) More than ogive

(i) Less than ogive

- First prepare less than cumulative frequencies in which case the frequencies are added serially from top to bottom. Then taking upper limits of the class intervals as the x-coordinate and the corresponding less than cumulative frequencies as the y-coordinates , points are plotted . The points thus obtained , joined freely will give the less than ogive . Obviously

  "less than ogive" is an increasing curve slopping upward from left to right and the slope of an elongated S .

Website :- https://www.arjun00.com.np

(ii) More than ogive

- First prepare more than cumulative frequencies in which case frequencies are added serially from bottom to the top. Then taking lower limits of the class intervals as x-coordinates and the corresponding more than cumulative frequencies as y-coordinates , points are plotted. The points thus obtained are joined freely will give the "more than . ogive" . It is a decreasing curve and slopes downwards from left to right and the shape of the elongated S upwards down .
- **Note:** If both less than and more than ogive are drawn on the same paper, they intersect at a point. The foot of the perpendicular from their point of intersection on the x-axis gives the value of the median .

Example:

- Draw both "less than" and "more than" ogives from the following data and compute the value of the median .

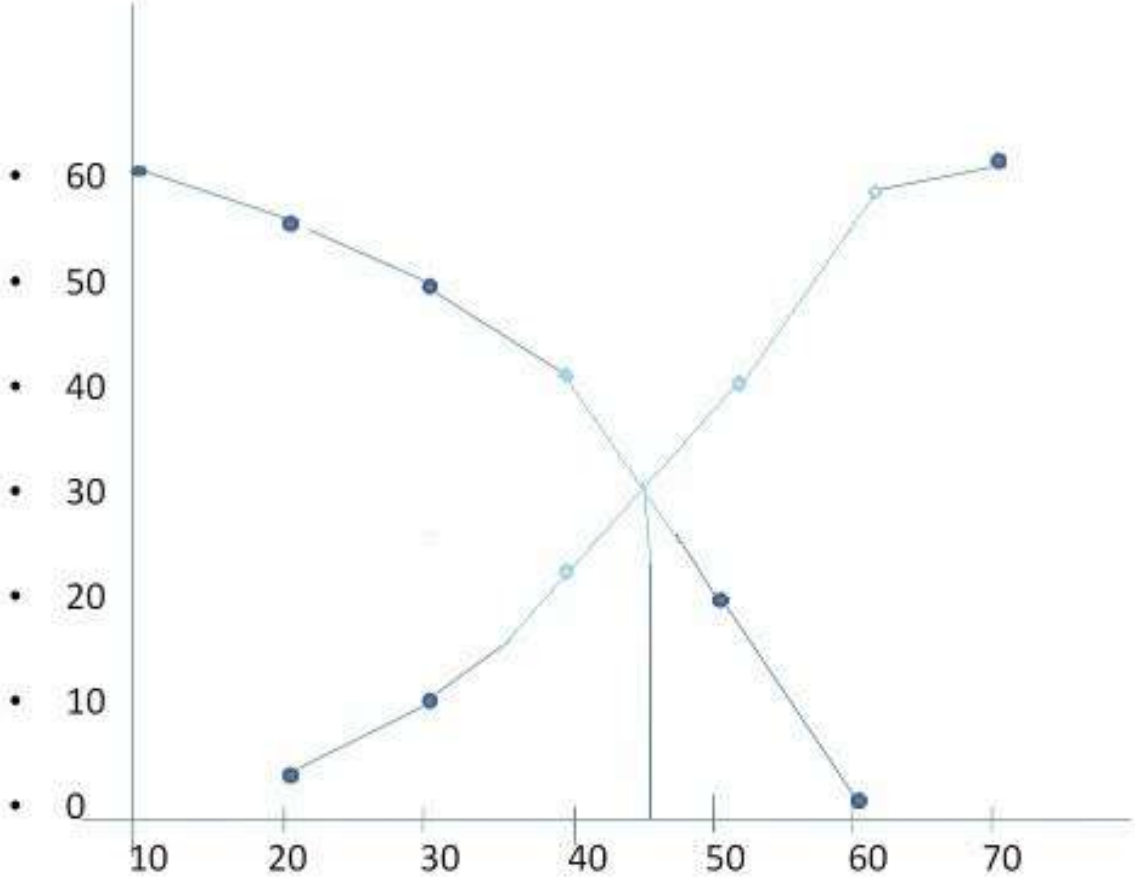
Rain fall (mm)	10-20	20-30	30-40	40-50	50-60	60-70
No, of days	4	6	10	20	18	2

Solution: first prepare less than and more than cumulative frequencies (c. f.)

For Less Than Ogive		For More Then Ogive	
Rain fall	Less than cf	Rain fall	More than cf
Less than 20	4	More than 10	56+4=60
Less than 30	4+6=10	More than 20	50+6=56
Less than 40	10+10=20	More than 30	40+10=50
Less than 50	20+20=40	More than 40	20+20=40
Less than 60	40+18=58	More than 50	2+18=20
Less than 70	58+2=60	More than 60	2

Website :- https://www.arjun00.com.np

"Less than" and "More than" Ogive



- From the graph , the value of the median = 45

Exercise 4

Diagrammatic and Graphical Representation

Website :- https://www.arjun00.com.np

Exercise 4

What are graphs and diagrams ? What are the importance and limitations of diagrammetic and graphical representation of statistical data ?

Distinquish between diagrammetic and graphical representation of data and explain the various principles of constructing diagrams and graphs.

Point out the general rules that should be followed in constructing diagrams.

Distinquish between

a) Histogram and Bar diagram

b) Pie-diagram and sub-divided bar diagram

c) Less than and more than ogives

What do you mean by diagrams ? Name the different types of diagrams commonly used and mention the situations where the use of each type of diagram would be appropriate.

6. What is meant by (a) Histogram (b) Frequency polygon (c) Frequency curve (d) ogive ? Explain their construction with the help of sketch.

7. Explain the different types of graphs used for representing a frequency distribution.

8. Write notes on (a) Histogram (b) Ogives (c) multiple bar diagram.

9. Draw the simple bar diagrams of the following data:

(i)

Dev. Reg.	Eastern	Central	Western	Mid-western	Far-western
Total length of Roads (in km):	1568	2344	1180	708	506

(ii)

Country:	Bangla -desh	Bhutan	India	Mal-dives	Nepal	Pakis-tan	Sri Lanka
Rural pop%	87	95	75	74	92	72	78

Website :- https://www.arjun00.com.np

10. (i) The table given below gives the density of population of two towns A and B during 3 census periods. Represent it by sub divided bar diagram:

Year	Town A	Town B
1931.	418	636
1941	470	700
1951	544	782

(ii) The data presented below gives the yearly profits (in "000" Rs.) of two companies A and B

	Profits (in "000" Rs)	
Year	Company A	Company B
1976-77	120	90
1977-78	135	95
1978-79	140	108
1979-80	160	120

Represent the above data by sub-divided bar diagrams.

11. Draw the multiple bar diagram of the following data:

	Production (in nearest "0000" m. tonne)		
Dev. Region	Maize	Wheat	Potato
Eastern	21	11	14
Central	27	25	14
Western	19	15	5
Middle-western	14	12	4
Far-western	6	8	3

Website :- https://www.arjun00.com.np

12. Represent the following data relating to the value of imports and exports from 1982 to 1985 by a multiple bar diagram.

Year	Export (Lakhs Rs.)	Import (Lakhs Rs.)
1982	610	624
1983	955	742
1984	660	578
1985	565	527

13. Represent the following percentage composition of land use by percentage bar diagram.

Type of Land use	Percentage Composition
Agriculture	18
Forest	37.6
Snow	15.3
Pasture	13.4
Water	2.7
Settlement and Roads	0.7
Others	12.3

14. (i) The following table shows the area in millions of square kms. of Ocean of the world.

Ocean:	Pacific	Atlantic	Indian	Antartic	Artic
Area:	70.8	41.2	28.5	7.6	4.8

Represent the data by a pie- diagram.

(ii) Draw a pie-diagram to represent the following population in a town.

Males	Females	Girls	Boys	Total
2000	1800	4200	2000	10000

15. Represent the following data by suitable diagram:

Campus	Humanities	Science	Commerce	Agriculture	Total
A	1200	800	600	400	3000
B	750	500	300	450	2000

16. The following table shows the result of the B. Com. examination of a university for past three years

	No. of Students Passed			
Year	I div.	II div.	III div.	failed
1967	48	84	162	200
1968	36	107	173	194
1969	22	121	160	324

Represent the above data by suitable diagram.

17. Draw the suitable diagrams for the following data:

Expenditure on	Family A Income Rs. 1000	Family B Income Rs. 6000
Food	300	300
Clothing	250	120
Education	50	100
Onforseen	380	140
Saving or deficit	[illegible] 20	60

Website :- https://www.arjun00.com.np

18. The following data relate to the expenditures of families per month:

Items of Expenditure	Family A	Family B
Food	200	300
Rent	100	200
Clothing	100	150
Education	50	200
Lighting	25	50
Miscellaneous	25	100

Represent these data by angular diagrams.

19. Represent the following data by means of a histogram:

Marks	Frequency	Marks	Frequency
0-10	4	40-50	22
10-20	10	50-60	18
20-30	16	60-70	8
30-40	22	70-80	2

20. The following are the temperature records of 60 days.

Temp (°c):	5-10	10-15	15-20	20-25	25-30	30-35
No. of days:	6	12	15	17	8	2

Draw the histogram, frequency polygon and frequency curve of the above data

21. Draw the histogram from the following data:

Mid value:	45	55	65	75	85	95	105	115
Frequency:	5	7	11	15	8	4	3	2

22. Draw less than Ogive from the following data. Find the number of workers earning less than Rs. 55 per week. Also find the median wage.

Weekly wages (Rs.):	0-20	20-40	40-60	60-80	80-100	100-120
No. of workers:	40	51	64	38	7	5

Obtain the number of workers whose wages lie between Rs. 50 and Rs.

Website :- https://www.arjun00.com.np

23. Draw the less than Ogive and estimate the value of the median on the basis of the data given below:

Mid value:	18	25	32	39	46	53	60
Frequency:	10	15	32	42	26	12	9

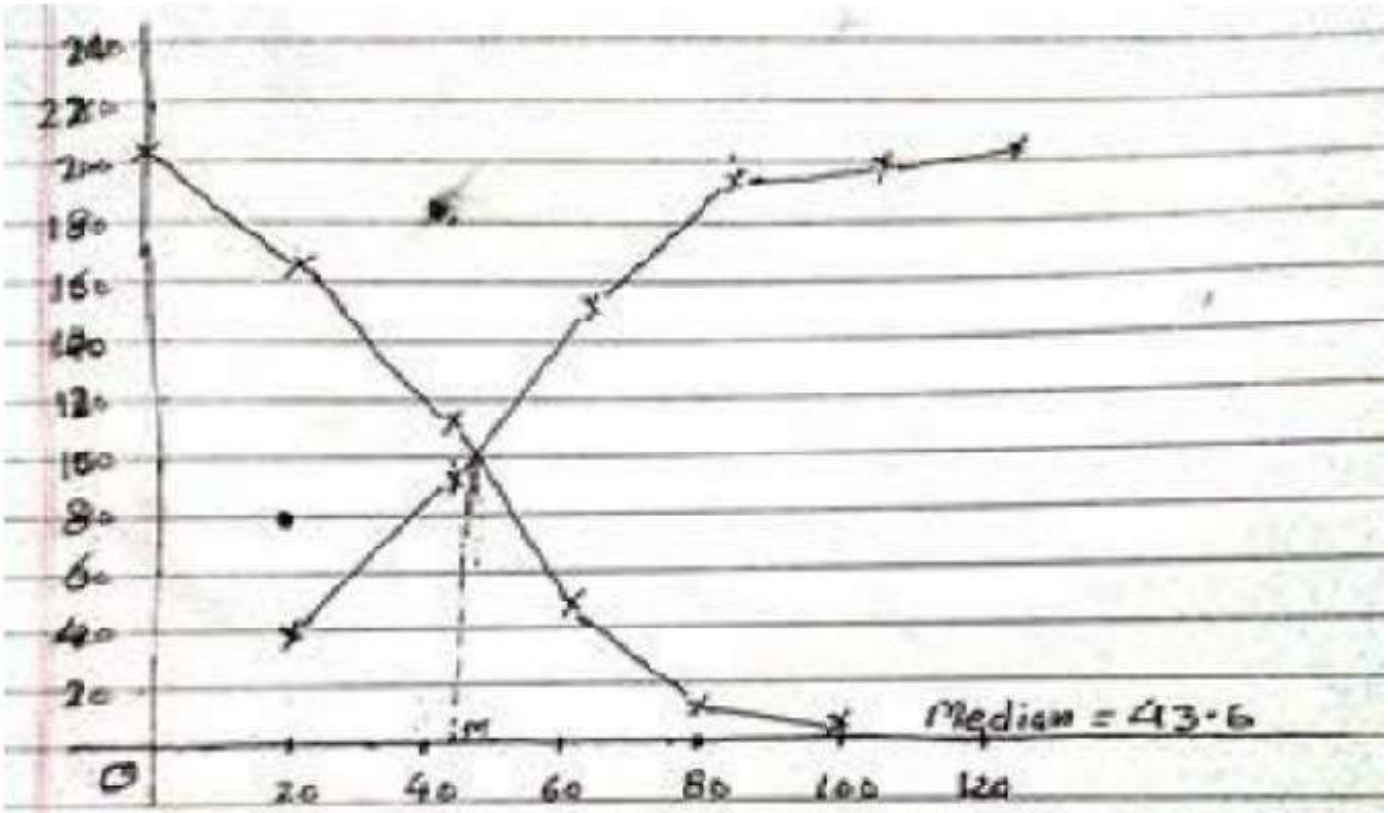
24. Draw the "less than" and "more than" ogive from the data given below:

Marks:	10-20	20-30	30-40	40-50	50-60	60-70
No of students:	8	12	20	6	4	10

Find the median.

22. solution:

For less than ogive than		For more than ogive	
Weekly wages	Less than cf	Weekly wages	More than cf
Less than 20	40	More than 0	165+40=205
Less than 40	40+51 = 91	More than 20	114+51=165
Less than 60	91+64 = 155	More than 40	50+64=114
Less than 80	155+38 = 193	More than 60	12+38=50
Less than 100	193+7 = 200	More than 80	5+7=12
Less than 120	200+5 = 205	More than 100	5



No of workers whose wages less than Rs 55

$= 40 + 51 + \frac{15}{20} \times 64 = 91 + 48 = 139$

No of workers whose wage lie between Rs 50 and Rs 90

Website :- https://www.arjun00.com.np

Unit 5

Summarizing Data Set

Introduction

- After the collection of data , the next step is to analyze it , since huge masses of data are confusing and difficult to remember, so we need a unique value representing them. The averages are the measures which condence a huge mass of data into single value representing the whole data. Averages are the typical values around which most of the data tend to cluster. These are the values which lie between two extreme observations of the entire data and give us the idea about the concentration of the value in the central part of the distribution. Measure of such single value is known as measure of central tendency .

Measure of Central Tendency

- The following are the measure of central tendency which are commonly used as:
- (1) Mean
- (2) Median
- (3) Mode

Website :- https://www.arjun00.com.np

(1) Mean

- Mean is a measure of central value of given set of observation. It has three types:
- (a) Arithematic mean
- (b) Geometric mean
- (c) Harmonic mean
- **(a) Arithematic mean:**

  The arithematic mean of a given set of observation is their sum divided by the number of observations. The arithematic mean is generally denoted by $\bar{x}$

Thus if $x_1, \bar{x}...\bar{x}_n$ is the given set of observations, then their arithematic mean is given by

$$\bar{x} = \frac{x_1 \pm \bar{x} + \cdots \bar{x}_n}{n} = \frac{\sum x}{n}$$, where $\sum x$ = the sum of observations and n = the number of observations .

Individual Series

Let $x_1, x_2 \ldots x_n$ be n individual observations , then their mean is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n} = \frac{\sum x}{n}$$. This is direct method.

Website :- https://www.arjun00.com.np

Short cut method

When the number of observation be too large and the number of the variate value is very large, then we use short cut method by the formula

$$\bar{x} = a + \frac{\sum d}{n} \quad ,$$

Where $d = x - a = deviation\ and\ \ a = \ assumed\ mean$

Example:1 The expenditure of 6 families in rupees are given

Family	A	B	C	D	E	F
Exp.(Rs.)	300	325	375	400	425	500

Calculate the arithmetic mean by (a) direct method (b) short cut method.

Solution:

Family	Expd (x)	d = x – a(375)
A	300	-75
B	325	-50
C	375	0
D	400	25
E	425	50
F	500	125
	$\sum x = 2325$	$\sum d = 75$

Website :- https://www.arjun00.com.np

(a) Direct method: $\bar{x} = \frac{\sum x}{n} = \frac{2325}{6} = 387.5$

(b) Short cut method : $\bar{x} = a + \frac{\sum d}{n} = 375 + \frac{75}{6} = 387.5$

Discrete series:

Let $f_1, f_2, f_3, \ldots f_n$ be the frequencies of the variate values $x_1, x_2, x_3, \ldots x_n$ respectively , then their arithematic mean is given by $\bar{x} = \frac{f_1x_1+f_2x_2\ldots+f_nx_n}{f_1+f_2+\cdots..f_n} = \frac{\sum fx}{\sum f} = \frac{\sum fx}{N}$

Where $N = \sum f$ =total frequency . This is direct method .

Short cut method:

$$\bar{x} = a + \frac{\sum fd}{\sum f} = a + \frac{\sum fd}{N}$$

Where a = assumed mean

d = x - a = deviation of the item taken from assumed mean .

Continuous series: ***Step deviation method***

$$\bar{x} = a + \frac{\sum fd'}{N} \times h$$

where $d' = \frac{x-a}{h}$ and $h = class\ size$

Website :- https://www.arjun00.com.np

EXAMPLE 2:

Find the arithmetic mean from the data given below by using

(i) Direct method (ii) Step-deviation method.

Class boundaries:	15-25	25-35	35-45	45-55	55-65	65-75
Frequency (f):	4	11	19	14	0	2

Solution

COMPUTATION OF ARITHMETIC MEAN

Class boundaries	Mid. value(x)	f	fx	$\frac{x-40}{10} = d'$	fd'
15-25	20	4	80	-2	-8
25-35	30	11	330	-1	-11
35-45	40	19	760	0	0
45-55	50	14	700	1	14
55-65	60	0	0	2	0
65-75	70	2	70	3	6
		N =50	$\Sigma fx = 2010$		$\Sigma fd' = 1$

(a) Direct method:

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{2010}{50} = 40.2$$

(b) Step deviation method:

$$\bar{x} = a + \frac{\Sigma fd'}{N} \times h = 40 + \frac{1}{50} \times 10 = 40.2$$

Website :- https://www.arjun00.com.np

5.4.4 Weighted arithmetic average

The arithmetic average discussed above is the simple arithmetic average in which all the items are assumed to be equally important in the distribution. But in practice, this may not be so. The importance of some items in a distribution may be greater than the other, so in such cases proper weightage should be given to various items. Now, we define the following weighted arithmetic average in which proper weight is considered.

Let $w_1, w_2, w_3, \ldots\ldots; w_n$ be the weights given to the variate values $x_1, x_2, x_3, \ldots\ldots x_n$ respectively, then their weighted arithmetic mean denoted by $\overline{x}_w$ is defined by

$$\overline{x}_w = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots w_n}$$

$$\overline{x}_w = \frac{\sum wx}{\sum w} \ldots\ldots\ldots (iv)$$

5.4.5 Properties of arithmetic mean

The following are the important properties possessed by the arithmetic Mean.

1) The algebraic sum of the deviations of the items taken from their arithmetic mean is zero.

Let x be the variate values of the frequency distribution and $\overline{x}$ their arithmetic mean. Then, the sum of the deviations of the items from the arithmetic mean

$$= \sum f(x - \overline{x})$$

$$= \sum fx - \overline{x}\sum f = N\overline{x} - N\overline{x} = 0$$

Website :- https://www.arjun00.com.np

2. The sum of the squares of the deviations of the items is minimum when taken from the arithematic mean .

Let 'a' be an arbitrary number from which deviations are to be taken and S be the sum of the squares of the deviations taken from 'a' , then

$$S = \sum f(x-a)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (i)$$

Differentiating both sides w.r.to 'a'

$$\frac{dS}{da} = -2\sum f(x-a) \ldots\ldots\ldots\ldots\ldots (ii)$$

Again, differentiating both sides w.r.to 'a'

$$\frac{d^2S}{da^2} = 2\sum f = 2N \ldots\ldots\ldots\ldots\ldots\ldots\ldots (iii)$$

For maxima or minima , $\frac{dS}{da} = 0$

Or $-2\sum f(x-a) = 0$ Or $\sum f(x-a) = 0$

Or $\sum fx - \sum fa = 0$ Or $\sum fx = \sum fa$

Or $\sum fx = a\sum f$ Or $\sum fx = aN$

Or $a = \frac{\sum fx}{N} = \bar{x}$ which is arithematic mean.

Put $a = \bar{x}$ in (iii) we get

$\frac{d^2S}{da^2} = 2N$ which is positive

Hence S has minimum value at $a = \bar{x}$. Proved.

Website :- https://www.arjun00.com.np

3. Combine Mean

If n_1 *and* n_2 be the sizes and x_1 *and* $\bar{x}_2$ be the arithematic mean of two component series , then the combined mean is denoted by x_{12} is given by

$$\bar{x}_{12} = \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$$

5.4.6 Merits and Demerits of arithmetic mean

Merits:

i) It is rigidly defined.
ii) It is based on all observations.
iii) It is simple to understand and easy to calculate.
iv) It is suitable for further mathematical treatments.
v) It is least affected by fluctuation of sampling.

Demerits :

i) It is very much affected by extreme observations.
ii) It cannot be computed accurately in case of open end classes.
iii) It gives sometimes fallacious conclusion.
iv) It cannot be determined by inspection or by graphical method.
v) It cannot be used if we are dealing with qualitative characteristics which cannot be measured quantitatively.

Website :- https://www.arjun00.com.np

Example:3 In class of 20 students , the average of 12 boys is 15 years and the average age of 8 girls is 17 years . Find the average age of the class .

Solution: No. of boys $(n_1) = 12$, average age of boys $(x_1) = 15$

No. of girls $(n_2) = 8$, average age of girls $(\bar{x}_2) = 17$

Combined mean age $(x_{12}) = ?$

$$\therefore x_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{12\times15+8\times17}{12+8} = \frac{316}{20} = 15.8 \text{ years.}$$

Geometric Mean

The geometric mean of the n non-zero and non-negative variate values is the nth root of their product .

For Individual series

Let $x_1, x_2, x_3, \ldots \ldots x_n$ be n non-zero and non-negative variate values ,then their geometric mean denoted by G is defined by

$$G = \sqrt[n]{x_1 . x_2 . x_3 \ldots \ldots . x_n}$$

Or $$G = antilog\left[\frac{\sum \log x}{n}\right]$$

For Discrete and continuos Series

If $f_1, f_2, f_3, \ldots \ldots f$ be the corresponding frequencies of n non-zero and non-negative variate values $x_1, x_2, x_3, \ldots \ldots x_n$, then their geometric mean is given by

$$G = antilog\left[\frac{\sum f \log x}{N}\right]$$, where N is the total frequency .

Example:4 Calculate the geometric mean of the following distribution:

Class	2-4	4-6	6-8	8-10
Frequncy	20	40	30	10

Solution:

Class	Mid-value(x)	Frequency(f)	log x	f.log x
2-4	3	20	0.4771	9.542
4-6	5	40	0.6990	27.960
6-8	7	30	0.8451	25.353
8-10	9	10	0.9552	9.442
		N=100		$\sum f.\log x = 72.395$

$$G = antilog\left[\frac{\sum f \log x}{N}\right] = antilog\left[\frac{72.395}{100}\right] = antilog(0.72495) = 5.297$$

Website :- https://www.arjun00.com.np

5.5.4 Merits and Demerits of G. M.

Merits:

i) It is rigidly defined.

ii) It is based on all the observations;

iii) It is suitable for further mathematical treatment.

iv) It is not affected much by fluctuation of sampling.

v) It is not most appropriate average while dealing with ratios and percentages.

Demerits:

i) It is difficult to understand.

ii) It is not easy to calculate.

iii) Geometric mean cannot be calculated if the variate value is zero or negative.

iv) It gives more weight to smaller item and less to the greater.

5.6 Harmonic mean

Harmonic mean is defined as the reciprocal of the arithmetic average of the reciprocals of the set of non-zero variate values.

5.6.1 Individual series

Let $x_1, x_2, x_3, \ldots\ldots, x_n$ ben non-zero variate values of the variable x, then their harmonic mean (H. M.) denoted by H is defined by

$$H = \frac{n}{\Sigma \frac{1}{x}} \qquad \ldots\ldots (x)$$

where n is the number of observations.

Website :- https://www.arjun00.com.np

5.6.2 Discrete series

If the n non-zero variate values $x_1, x_2, x_3, \ldots\ldots, x_n$ repeat $f_1, f_2, f_3, \ldots\ldots f_n$ times respectively then their harmonic mean H is defined by

$$H = \frac{N}{\sum f \times \frac{1}{x}} \qquad \ldots\ldots (xi)$$

where N = total frequency.

5.6.3 Continuous series

For a continuous series, the same formula (xi) i.e. $H = \frac{N}{\sum f \times \frac{1}{x}}$ is used to find the harmonic mean. In this case x is the mid values of the classes.

EXAMPLE 4.

Calculate the harmonic mean of the following distribution.

x:	3	5	7	9
f:	20	40	30	10

Solution:

COMPUTATION OF HARMONIC MEAN

x	f	$\frac{1}{x}$	$f \times \frac{1}{x}$
3	20	0.3333	6.666
5	40	0.2	8.000
7	30	0.1428	4.284
9	10	0.1111	1.111
	N = 100		$\sum f \times \frac{1}{x} = 20.061$

$$H = \frac{N}{\sum f \times \frac{1}{x}} = \frac{100}{20.061} = 4.985$$

Website :- https://www.arjun00.com.np

5.6.4 Merits and Demerits of H. M:

Merits:

i) It is based on all the observations.

ii) It is not much affected by fluctuation of sampling.

iii) It is the most appropriate average while dealing with speed.

Demerits:

i) It is not easy to understand.

ii) It is difficult to compute.

iii) It can not be determined when the variate value is zero.

iv) It gives greater importance to the smaller item.

5.7 Median

The variate value dividing the total number of observations into two equal parts is said to be the median. The median divides the whole observations into two halves, on half comprising all the values greater and the second half, all values less than median. It is the positional average

i.e. its value depends on the position occupied by a value in the frequency distribution. It is denoted by M_d Determination of median depends upon the series given.

5.7.1. Individual series

First we arrange the set of observations according to their according or descending order. If the number of observations is odd, the middle value gives the median. Again if the number of observations is even, there will be two middle values so the arithmetic average of two middle values gives the median. The formula for computing the median in case of individual series is

$$\text{Median} = \text{value of } \left(\frac{n+1}{2}\right)^{th} \text{ item} \qquad \ldots(5.6)$$

where n is the no of observations.

Website :- https://www.arjun00.com.np

EXAMPLE 5:

Find the median from the following set of observations.

5, 3, 2, 4, 10, 7, 12.

Solution:

Arranging the given set of observations according to their ascending order, we have

2, 3, 4, 5, 7, 10, 12

n = no. of observations = 7.

$$\text{Median} = \text{value of } \left[\frac{n+1}{2}\right]^{th} \text{ item}$$

$$= \text{value of } \left[\frac{7+1}{2}\right]^{th} \text{ item}$$

$$= \text{value of } (4)^{th} \text{ item} = 5$$

5.7.2 Discrete series

In finding the median in case of discrete series, we use the following procedure:

i) Arrange the data according to their ascending order of their magnitudes.

ii) Form the cumulative frequency table.

Website :- https://www.arjun00.com.np

iii) Use the formula, Median = value of $\left[\frac{N+1}{2}\right]^{th}$ item ..... (xiii)

where N = total frequency.

iv) Now see the cumulative frequency column and note the value corresponding to the cumulative frequency either equal to or greater than $\frac{N+1}{2}$. This gives the value of the mediam.

5.7.3 Continuous series

The following steps are to be used in finding the median in case of continuous series:

i) Prepare less than cumulative frequency distribution.

ii) Find N/2

iii) See cumulative frequency equal to or just greater than N/2 and note the class.

iv) The corresponding class contains the median value and is called the median class.

v) The value of the median is now computed by the formula

$$\text{Median} = l + \frac{\frac{N}{2} - c.f.}{f} \times h \text{ ..... (xiv)}$$

where l = lower limit of the median class

N = total frequency

c.f. = cumulative frequency preceding the median class

f = frequency of the median class

h = Size of the median class

Website :- https://www.arjun00.com.np

EXAMPLE 6.

Find the median from the following distribution:

Consumption of electricity (units)	20-30	30-40	40-50	50-60	60-70	70-80	89-90	90-100
No. of families	10	20	10	25	10	50	40	30

Solution:

COMPUTATION OF MEDIAN

Consumption of elect.	Frequency	Cumulative frequency (c.f.)
20-30	10	10
30-40	20	30
40-50	10	40
50-60	25	65
60-70	10	75
70-80	50	125
80-90	40	165
90-100	30	195

$$\frac{N}{2} = \frac{195}{2} = 97.5$$

∴ Median lies in the class 70-80.

Here, $l = 70$, $N/2 = 97.5$, $c.f. = 75$, $f = 50$,

$h = 80 - 70 = 10$, Median = ?

Now, $$\text{Median} = l + \frac{\frac{N}{2} - c.f.}{f} \times h$$

$$= 70 + \frac{97.5 - 75}{50} + 10$$

$$= 70 + 4.5 = 74.5$$

∴ Median = 74.5 units.

Website :- https://www.arjun00.com.np

5.7.4 Merits and Demerits of Median

Merits:

i) Median is rigidly defined.

ii) It is simple to understand and easy to calculate.

iii) Median is not affected by extreme observations.

iv) Median can be computed even for open-end classes.

v) Median can sometimes be located by inspection.

vi) Median can be obtained graphically.

vii) Median is only the average to be used while dealing with qualitative characteristics such as intelligence, beauty etc.

Demerits:

i) Arrangement of data according to magnitude is necessary.

ii) Median is not based on all observations.

iii) For an ungrouped data, if the number of observation is even, median cannot be determined exactly.

iv) Median is not suitable for further mathematical treatment.

v) For a small size sample, median is affected by fluctuation of sampling.

Website :- https://www.arjun00.com.np

Unit-5

Exercise 5

1. What are the various measures of central tendency and to what purposes are they used ?
2. What do you mean by an "average" ? What are the desirable properties for good average to possess ?
3. What do the "measures of central tendency" mean ? Describe briefly various methods of measuring central tendency. Point out relative merits and limitations of each of them.
4. Define arithmetic average, write the properties of an arithmetic average. Why is arithmetic average most commonly used amongst them ?
5. Under what circumstances would you consider appropriate to use arithmetic mean, geometric mean, harmonic mean, median and mode ?
6. What is an average ? Prove that the sum of the squares of the deviations of the items is minimum when taken from the arithmetic mean.
7. Write notes on:
   i) Partition values
   ii) Relation between different averages
   iii) Requisites of a good average
   iv) Choice of an average.
8. The following are the prices of shares of a company from Monday to Saturday:

Day :	M	T	W	TH	F	S
Price (Rs.):	200	210	208	160	220	250

Compute the average price by

(i) Direct method (ii) Short cut method.

9. Find the arithmetic mean from the following frequency table:

Height (in cms):	40	45	50	52	60	70
No. of Plants :	7	8	9	6	4	2

Also, calculate the mean by short cut method.

Website :- https://www.arjun00.com.np

10. Given that A. M. is equal to 7.3, find the missing frequency in the following data:

x:	5	6	7	8	9
f:	4	6	12	?	8

11. An examination was held to decide about the award of a scholarship in T. U. the weights of various subjects were different. The marks obtained by 3 candidates (out of 100 in each subject) are given below:

Subject	Weight	Marks A	Marks B	Marks C
Statistics	5	70	80	85
Mathematics	3	90	75	75
Physics	1	50	60	45
Chemistry	2	50	45	65

If the candidate getting the highest mark is to be awarded the scholarship, who should get it ?

12. Compute the arithmetic mean from the following distribution by

(i) Direct (ii) Short cut (iii) Step deviation methods.

Production:	0-10	10-20	20-30	30-40	40-50
No. of factories :	5	8	15	16	6

13. The marks scored by 60 students in an examination in statistics are given below. Form a frequency distribution with class-interval 10 and calculate the arithmetic mean by direct as well as step deviation method.

6	10	58	56	0	25	32	35	35	9	78	17	60	50	35
38	30	10	48	5	68	48	35	31	41	23	23	50	72	19
25	35	40	46	42	45	25	60	41	35	36	38	35	33	46
28	31	35	42	46	38	39	45	48	50	28	29	31	55	30

14. From the following data of income distribution, calculate the arithmetic mean. It is given that (i) the total income of the persons in the highest group is R. 435 and (ii) none is earning less than Rs 20.

Income (Rs)Below :	30	40	50	60	70	80	80 and over
No. of persons:	16	36	61	76	87	95	5

Website :- https://www.arjun00.com.np

15. From the information given below, find

(i) which factory pays larger amount as daily wage ?

(ii) what is the average daily wage for the workers of the two factories ?

	Factory A	Factory B
No. of wage earners:	250	200
Average daily wage:	Rs. 20	Rs. 25

16. Calculate the arithmetic mean, geometric mean and harmonic mean of the following observations:

32, 35, 36, 37, 39, 41, 43

and show that A. M. > G M. > H. M.

17. Find the G. M. and H. M. of the following data.

Variable:	75-125	125-175	175-225	225-275	275-325
Frequency:	10	20	30	6	4

18. Calculate the mean, median and mode from the following data of heights in cms of a group of students.

161, 162, 163, 161, 163, 164, 164, 160, 165, 163, 164, 165, 166, 164. Now suppose that a group of students whose heights are 160, 166, 159, 168, 167 and 170 cms. is added to original group. Find the mean, median, mode of the combined group.

19. The following marks have been obtained in three papers of statistics in examination by 12 students, in which paper is the general level of knowledge of the students highest ? Give reason.

A:	36	56	41	46	54	59	55	51	52	44	37	59
B:	58	54	21	51	59	46	65	31	68	41	70	36
C:	65	55	26	40	30	74	45	29	85	32	80	39

20. Determine the median and mode from the following data:

x:	20	25	30	35	40	45	50	55	60	70
f:	5	6	8	10	12	3	2	6	1	4

Website :- https://www.arjun00.com.np

21. Amend the following table and locate the median from the amended data:

Size	Frequency	Size	Frequency
10-16	10	30-35	28
16-17.5	15	35-40	30
17.5-20	17	40 and onwards	40
20-30	25		

22. The following table shows the age distribution of persons in a particular region:

Age (yrs) Below:	10	20	30	40	50	60	70	80
No. of persons ('000'):	2	5	9	12	14	15	15.5	15.6

Find the median age, why is the median a more suitable than the mean in this case ?

23. Calculate the median, quartiles, 6th decile and 70th percentile from the following data:

Marks less than:	10	20	30	40	50	60	70	80
No of students:	5	13	20	32	60	80	90	100

24. Find the median and mode from the following data:

Size	Frequency	Size	Frequency
over 5	320	over 25	105
over 10	295	over 30	50
over 15	250	over 35	20
over 20	190	up to 40	320

25. The table given below presents the frequency distribution of weights of 80 apples selected at random from a big consignment.

Wt. (in gms)	Frequency	Wt. (in gms)	Frequency
110-119	5	150-159	16
120-129	7	160-169	10
130-139	12	170–179	7
140-149	20	180-189	3

Find the median, two quartiles, 4th decile and the 80th percentile.

26. Monthly income of 400 employees in an organization are given below:

Website :- https://www.arjun00.com.np

Monthly income	Employees
Below Rs. 80	10
" Rs. 120	35
" Rs. 160	165
" Rs. 200	310
" Rs. 240	380
" Rs. 280	400

Obtain the limits of income of central 60% of observed employees.

27. (i) In a moderately skewed distribution, if arithmetic mean = 24.6 and the mode = 26.1, find the value of the median.
(ii) In moderately asymmetrical distribution, the value of mode and median are 20 and 24 respectively. Locate the value of mean.

28. Calculate A.M., median and mode from the following data:

Value:	0-9	10-19	20-29	30-39	40-49	50-59	60-69
Frequency:	8	10	15	20	15	10	8

Explain the nature of distribution also.

29. Find the simple and weighted arithmetic mean of the first n natural numbers, the weights being the corresponding numbers.

30. A variate takes values a, ar, ar^2 ....., ar^{n-1} each with frequency unity, Show that the A. M. 'A' is $\frac{a(1-r^n)}{n(1-r)}$ the G. M. 'G' is and $ar^{(n-1)/2}$ the H. M. 'H' is $\frac{an(1-r)r^{n-1}}{1-r^n}$. Prove that A. H. = G^2

Answers

(8) Rs. 208 (9) 49.5 cms (10) 10 (11) C (12) 27
(13) 37 marks (14) Rs. 48 (15). (i) Same (ii) Rs. 22.22
(16) A. M.=37.57, G. M. = 37.41, H. M. = 37.25
(17) G. M. = 174.1 H.M = 166.40
(18) Original group: Mean = 163.2, M_d = 163.5, M_o = 164,
Combined group: Mean = 163.75, M_d = 164. M_o = 164.
(19) B. (20) 35, 35 (21) 32.67 (22) 27
(23) 34.17, 46.43, 57.5, 50, 55 (24) 21.76, 22.27
(25) 147.5 gms, 136.17 gms, 159.5 gms., 143.5 gms, 163.5 gms,
(26) 133.85-205. 71 (27) (i) 25.1 (ii) 26.

28 Mean = M_d = M_o = 34.5, Symmetrical. 29. $\frac{n+1}{2}$, $\frac{2n+1}{3}$

Website :- https://www.arjun00.com.np

Chapter 6

MEASURES OF DISPERSION

6.1. Introduction and meaning

Averages give us the idea of concentration of the items around the central part of the distribution. But the averages only do not give the clear picture about the distribution because two distribution with same averages may differ in the scatterness of the items from the central value. For this, we see the following two series.

								$\bar{x}$	M_o	M_d
A:	25	26	27	27	27	28	29	27	27	27
B:	0	10	18	27	27	27	80	27	27	27

From the above table, we see that mean, mode and median of two series A and B are same. Only with these results we cannot say that the two series A and B are similar. Because, the difference of the items from the average in B is more in comparison to A. So in series A, items are concentrating more around the central value but the scatterness of the items from central value in series B is more. Hence, though two series A and B have same averages, they can not be said similar because they are differently constituted.

Actually the meaning of dispersion is the scatterness of the items from the central value. So, dispersion is defined as the measure of variation in the items from the central value;.

The main objects of measuring variability are as follows.

i) To determine the reliability of an average.

ii) In devising a system of quality control.

iii) To compare two or more series with regard to their variability.

iv) To help in using other statistical tools.

6.2 Requisites of a good measure of dispersion

The following are the necessary characteristics to be satisfied by different measures of dispersion to become a good measure.

i) It should be rigidly defined.

Website :- https://www.arjun00.com.np

ii) It should be simple to understand and easy to calculate.
iii) It should be based on all observations.
iv) It should be amendable for further mathematical treatment.
v) It should be least affected by fluctuation of sampling.
vi) It should not be unduly affected by extreme values.

6.3 Absolute and Relative measures

Those measures of dispersion whose units are same as the units of the given series are known as the absolute measure of dispersion. These types of dispersions can be used only in comparing the variability of the series (or distribution) having the same units. Comparison of two distributions with different units can not be made with absolute measures. On the otherhand, the relative measures of dispersions are obtained as the ratio of absolute measure of dispersion to suitable average and are thus a pure number independent of units. Hence two distributions with different units can be compared with the help of relative measures of dispersion.

6.4 Methods of measuring dispersion

The following are the methods of measuring dispersion

i) Range
ii) Semi interquartile range or Quartile deviation
iii) Mean deviation or Average deviation
iv) Standard deviation

6.5 Range

The simplest method of studying the variation in the distribution is the range. Range is defined as the difference between the largest item and the smallest item in the set of observations. So in a set of observations if L is the largest item and S, the smallest item, then the range is defined by

$$\text{Range} = L - S.$$

In a grouped frequency distribution, range is the difference between the upper limit of the largest class and the lower limit of the smallest class. Some authors used to take the difference between the middle points of the largest and the smallest classes to get the range.

Range is the absolute measure of dispersion. It cannot be used to compare two distributions with different units. So, the relative measures corresponding to the range known as coefficient of range is defined by

$$\text{Coefficient of range} = \frac{L - S}{L + S}$$

Website :- https://www.arjun00.com.np

Ex. Find the range from the following data:

Day:	S	M	T	W	Th	F	Sat
Temp (°c):	32	34	31	25	21	35	36

Solution

L=Largest item = 36, S = Smallest item = 21

∴ Range = Largest item - Smallest item

$= 36 - 21 = 15°C$

6.5.1 Merits and Demerits of range

Merits:

i) It is rigidly defined.

ii) Range is simple to understand and easy to calculate

iii) Only minimum time is required to know the variability with the help of range.

Demerits:

i) It is not based on all observations.

ii) Range is affected by fluctuation of sampling.

iii) Range is affected by extreme values.

iv) Range cannot be calculated in case of open end classes.

v) Range is not suitable for further mathematical treatment.

6.6 Semi-interquartile range or Quartile deviation

The measure of dispersion depending upon the lower and upper quartiles is known as the Quartile deviation. The difference between the upper and lower quartiles is known as 'Interquartile range". Half the interquartile range is known as semi-interquartile range or quartile deviation.

$$\therefore \quad \text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

$$\text{Q.D.} = \frac{(Q_3 - M_d) + (M_d - Q_1)}{2}$$

So, quartile deviation gives the average amount by which the two quartiles differ from the median. Since in a distribution 25% of the observations will fall below Q_1 and 25% of the observations will lie

Website :- https://www.arjun00.com.np

above Q_3 (according to ascending order of magnitude), so 50% of the observations will lie between Q_1 and Q_3. Less the quartile deviation more will be the uniformity or less will be the variability in the central 50% of the items. Again greater the Q.D., less will be the uniformity or greater will be the variability in the central 50% of the items. Q.D. is the absolute measure of dispersion.

The relative measure based on the lower and upper quartiles known as coefficient of quartile deviation is given by

$$\text{Coeff of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

EXAMPLE 2:

From the following table giving the heights of students, calculate the quartile deviation and the coefficient of Q.D.

Height (cm) :	153	155	157	159	161	163
No of students:	25	21	28	20	18	20

Solution:

COMPUTATION OF Q.D.

Height (cms.)	Frequency	c.f.
153	25	25
155	21	46
157	28	74
159	20	94
161	18	112
163	24	136

$$\frac{N+1}{4} = \frac{(136+1)}{4} = 34.25$$

$$\frac{3(N+1)}{4} = 102.75$$

Q_1 = Size of (34.25)th item = 155

Q_3 = Size of (102.75)th item = 161

Now, quartile deviation $= \frac{Q_3 - Q_1}{2} = \frac{161 - 155}{2} = 3$ cms

and coeff. of Q.D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{161 - 155}{161 + 155} = 0.02$

Website :- https://www.arjun00.com.np

6.6.1 Merits and Demerits of Q.D.

Merits:

i) It is rigidly defined.

ii) It is simple to understand and easy to calculate.

iii) It is better measure of dispersion in comparison to range as it is based on 50% of central items.

iv) It is not affected by extreme values.

v) It can be calculated even when end classes are open.

Demerits:

i) It is not based on all observations.

ii) Q.D. is affected by fluctuation of sampling.

iii) Q.D. is not suitable for further mathematical treatment.

6.7 Mean deviation (Average deviation)

Range and Q.D. are not the better measure of dispersion as all the items are not included in both cases and they do not show the variations of the items from an average, thus ignor the formation of the distribution. Now mean deviation shows the variation of the items from average.

Mean deviation is defined as the arithmetic mean of the deviations of the items from mean, median or mode, when all deviations are considered positive.

If $\overline{x}$, M_d and M_o be the arithmetic mean, median and mode of the set of variate values x, then the mean deviations (M.D.) are computed by the following formulae:

$$\text{M.D. from mean} = \frac{\Sigma/x - \overline{x}/}{n} = \frac{\Sigma/d/}{n} \quad \ldots\ldots \text{(i)}$$

$$\text{Also, M.D. from mean} = \frac{\Sigma f/x - \overline{x}/}{N} = \frac{\Sigma f/d/}{N} \quad \ldots\ldots \text{(ii)}$$

where $/d/=/x-\overline{x}/$ and read as modulus of d or $x - \overline{x}$. n = number of observations and N = total frequency.

The formula (i) is for the individual series and (ii) for discrete as well as the continuous series. For continuous series x is the mid value of the classes.

M.D. from median and mode can similarly be obtained by replacing $\overline{x}$ by M_d and M_o respectively.

Website :- https://www.arjun00.com.np

There is an advantage of finding the mean deviations from the median as the sum of the deviations of the items from median is least when signs are ignored (proved below). However in practice mean deviation from the mean is mostly used.

The relative measure of mean deviation is defined as follows:

$$\text{Coefficient of M.D. from mean} = \frac{\text{M.D. from mean}}{\text{mean}}$$

$$\text{Coefficient of M.D. from median} = \frac{\text{M.D. from median}}{\text{median}}$$

In a discrete distribution, the mean deviation from the median is the least.

Let x_i be the variate value, N the total frequency and 'a' an arbitrary point. Then the mean deviation of the items from an arbitrary point 'a' denoted by S is given by

$$S = \frac{1}{N} \sum_{i=1}^{n} f_i /x_i - a/$$

we shall show that S will be minimum when a is the median.

Let $x_m > a > x_{m+1}$

which implies $x_i > a$ for $i = 1, 2, 3, \ldots\ldots, m$.

and $x_i < a$ for $i = m + 1, \ldots\ldots, n$.

$$\text{Now} \quad S = \frac{1}{N} \sum_{i=1}^{n} f_i \mid x_i - a \mid$$

$$= \frac{1}{N} \sum_{i=1}^{m} f_i (x_i - a) + \frac{1}{N} \sum_{i=m+1}^{n} f_i(a - x_i)$$

$$\frac{\partial S}{\partial a} = -\frac{1}{N} \sum_{i=1}^{m} f_i + \frac{1}{N} \sum_{i=m+1}^{n} f_i$$

For maxima or minima, $\partial s/\partial a = 0$

$$\text{or,} \quad -\frac{1}{N} \sum_{i=1}^{m} f_i + \frac{1}{N} \sum_{i=m+1}^{n} f_i = 0$$

Website :- https://www.arjun00.com.np

$$\therefore \quad \sum_{i=1}^{m} f_i = \sum_{i=m+1}^{n} f_i$$

which shows that a is the median.

Since $x_i > a$ for $i = 1, 2, \ldots\ldots, m$ and $x_i < a$ for $i = m + 1, \ldots\ldots n$ so for different values of 'a' greater than the median S is the decreasing function of 'a' and for different values of 'a' less than the median, S is the increasing function of 'a' and also $\partial S/\partial a = 0$, hence it follows that the mean deviation from the median is less than that measured from any other value.

6.7.1 Merits and Demerits of M.D.

Merits:

i) It is easy to understand and calculate.

ii) It is based on all observations.

iii) As the deviations are taken from the central values, so comparison of two distributions about their formation can easily be made.

Demerits:

i) The greatest drawback of mean deviation is that the algebraic signs are ingored.

ii) It is not capable of algebraic treatments.

iii) It cannot be computed in case of open end classes.

iv) It does not give the satisfactory result when deviation are taken from mode as mode is ill defined.

6.8 Standard deviation

The standard deviation is the absolute measure of dispersion in which the drawbacks present in other measure of dispersion are removed. It is said to be the best measure of dispersion as it satisfies most of the requisites of a good measure of dispersion.

Standard deviation (s.d.) is defined as the positive square root of the mean of the square of the deviations taken from the arithmetic mean. It is denoted by σ.

If x be the variate values and $\overline{x}$ their arithmetic mean then the s.d denoted by σ_x or simply σ is given by

Website :- https://www.arjun00.com.np

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} \ldots\ldots \text{(i)}$$

Also $$\sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{N}} = \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2}$$

where n = no. of observation, N = total frequency

The formula (i) is for individual series and (ii) for discrete as well continuous series, x is the mid. value of each class for continuous series.

When deviation are taken from the assumed mean, then the formula for s.d. take the following forms:

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} \ldots\ldots\ldots\ldots \text{(iii)}$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \ldots\ldots\ldots\ldots \text{(iv)}$$

and $$\sigma = h \times \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \ldots\ldots \text{(v)}$$

where $d = x - a$, $d' = \frac{x-a}{h}$, a = assumed mean, h = common factor.

The formula (iii) and (iv) are known as short cut method of finding s.d. for individual and discrete series respectively and (v) known as "step deviation method" is mainly used for continuous series.

6.8.1 Difference between mean deviation and standard deviation

The following are the main differences between the mean deviation and the standard deviation.

(i) Algebraic signs are ignored in case of mean deviation where as in standard deviation algebraic signs are also taken into account,

(ii) Deviations can be taken from mean, mode or median in case of M.D. but in s.d. only the deviations have to be taken from the mean.

(iii) Standard deviation is used in skewness, kurtosis, coefficient of correlations, regression analysis, normal distibution and sampling theory but M.D. is not.

Website :- https://www.arjun00.com.np

Below, we prove the following result relating to the standard deviation and mean deviation from mean.

1. **In a discrete distribution the standard deviation is not less than the mean deviation from the mean.**

Let x_i be a variate value and f_i its corresponding frequency. If $\overline{x}$ be the arithmetic mean, then s.d. (σ) and M.D. from mean are defined by

$$\sigma = \sqrt{\frac{\Sigma f_i (x_i - \overline{x})^2}{N}} \quad \text{and} \quad \text{M.D.} = \frac{\Sigma f_i \mid x_i - \overline{x} \mid}{N}$$

where N is the total frequency.

We have to prove that

$$\sigma \geq \text{M.D. from mean}$$

or, $\sigma^2 \geq (\text{M.D. from mean})^2$

or, $\frac{1}{N} \Sigma f_i (x_i - \overline{x})^2 \geq \left\{ \frac{1}{N} \Sigma f_i / x_i - \overline{x} / \right\}^2$

Put $x_i - \overline{x} = d_i$

Then, $N \Sigma f_i^2 d_i^2 \geq (\Sigma f_i / d_i /)^2$

or, $(f_1 + f_2 + \ldots\ldots + f_n)\ (f_1 d_1^2 + f_2 d_2^2 + \ldots + f_n d_n^2)$
$\geq (f_1 / d_1 / + f_2 / d_2 / + \ldots + f_n d_n)^2$

or, $f_1 f_2 (d^2_1 + d_2^2) + \ldots \geq 2 f_1 f_2 / d_1 / \ / d_2 / + \ldots$

or, $f_1 f_2 (/d_1/ - /d_2/)^2 + \ldots \geq 0$

which is always true. Hence the result follows.

6.8.2 Properties of the standard deviation

Now, we shall show that the **standard deviation is independent of origin but not of scale.**

Let x be the variate values and $\overline{x}$ their arithmetic mean. Then, s.d. denoted by σ_x is defined by $\overline{x}$

$$\sigma_x = \sqrt{\frac{\Sigma f (x - \overline{x})^2}{N}}$$

where N = total frequency

Website :- https://www.arjun00.com.np

Let us take 'a' as the new origin and scale be h times original scale so that

$$u = \frac{x - a}{h}, \qquad \text{then } x = a + hu$$

Also, $\bar{x} = a + h\bar{u}$

$$x - \bar{x} = a + hu - a - h\bar{u} = h(u - \bar{u})$$

Now, $\sigma_x = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{N}} = h\sqrt{\frac{\Sigma f(u - \bar{u})^2}{N}} = h\sigma_u$

$\therefore \quad \sigma_x = h\,\sigma_u$

where σ_u is the standard deviation for the variate u.

This result proves that the standard deviation is independent of origin but not of scale.

EXAMPLE 3:

Find the standard deviation of the following data:

100, 150, 200, 250, 300, 350, 400.

Solution

COMPUTATION OF S.D.

x	$x - \bar{x}$ (250)	$(x - \bar{x})^2$
100	– 150	22500
150	–100	10000
200	– 50	2500
250	0	0
300	50	2500
350	100	10000
400	150	22500
$\Sigma x = 1750$		$\Sigma(x - \bar{x})^2 = 70000$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1750}{7} = 250$$

Again, $\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} = \sqrt{\frac{70000}{7}} = 100$

Website :- https://www.arjun00.com.np

6.8.2 Merits and Demerits of Standard deviation

Merits:

i) It is rigidly defined.
ii) It is based on all the observations.
iii) It is least affected by fluctuation of sampling
iv) It is suitable for further mathematical treatment.

Demerits:

i) It is difficult to compute.
ii) It gives greater weight to the extreme values and less to those which are near to the mean.
iii) Standard deviation cannot be calculated for open end classes.

6.9 Variance and mean square deviation

The square of the standard deviation is known as the variance. It is denoted by σ^2 or μ_2. Variance is calculated by the following formula:

$$\sigma^2 = \frac{\Sigma f(x-\overline{x})^2}{N} = \frac{\Sigma fx^2}{N} - \left[\frac{\Sigma fx}{N}\right]^2$$

The mean square deviation denoted by s^2 is defined by

$$s^2 = \frac{\Sigma f(x-\overline{x})^2}{N}$$

where a is an arbitrary number.

So, in a mean square deviation, the deviations are taken from the arbitrary number 'a'.

The positive square root of the mean square deviation is known as the root mean square deviation. It is denoted by s. Thus,

$$s = \sqrt{\frac{\Sigma f(x-a)^2}{N}}$$

6.10 Relation between σ^2 and s^2

By the definition of variance and mean square deviation, we have

$$\sigma^2 = \frac{\Sigma f(x-\overline{x})^2}{N} \quad \text{and} \quad s^2 = \frac{\Sigma f(x-a)^2}{N}$$

Website :- https://www.arjun00.com.np

$$\text{Now, } \sigma^2 = \frac{1}{N}\Sigma f(x - \overline{x})^2 = \frac{1}{N}\Sigma f\left[(x - a) - (\overline{x} - a)\right]^2$$

$$= \frac{1}{N}\Sigma f(x - a)^2 - 2(\overline{x} - a)\frac{1}{N}\Sigma f((x - a) + (\overline{x} - a))^2$$

$$= \frac{1}{N}\Sigma f(x - a)^2 - 2(\overline{x} - a)^2 + (\overline{x} - a)^2$$

$$\sigma^2 = s^2 - (\overline{x} - a)^2$$

which is the required relation between σ^2 and s^2.

Cor 1: As $(x - a)^2$ is non-negative so $\sigma^2 \leq s^2$ i.et. he mean square deviation is not less than the variance or, **the root mean square deviation is not less then the standard deviation.**

Cor 2: The equality in $\sigma^2 \leq s^2$ holds when $\overline{x} - a = 0$, or $a = \overline{x}$ i.e. s^2 will be least when $a = x$.

Hence the standard deviation is the least possible of root mean square deviation.

6.11 Combined standard deviation

If n_1 and n_2 be the sizes, $\overline{x_1}$ and $\overline{x_2}$ the arithmetic means and σ_1 and σ_2 the respective standard deviations of two component series, then their combined standard deviation denoted by σ_{12} is given by

$$\sigma_{12} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2}{n_1 + n_2}}$$

where $d_1 = \overline{x}_1 - \overline{x}_{12}$, $d_2 = \overline{x}_2 - \overline{x}_{12}$

and $\overline{x}_{12}$ = combined mean $= \dfrac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}$

6.12 Coefficient of variation

Standard deviation is the absolute measure of dispersion. The relative measure of dispersion based on the standard deviation is known as the coefficient of standard deviation. Symbolically it is defined by

$$\text{Coefficient of s.d.} = \frac{\text{s.d.}}{\text{Mean}} = \frac{\sigma}{x}$$

Website :- https://www.arjun00.com.np

The coefficient of dispersion based on standard deviation multiplied by 100 is known as the Coefficient of Variation (C.V.) If $\bar{x}$ be the arithmetic mean and σ, the standard deviation of the distribution, then C.V. is defined by

$$\therefore C.V. = \frac{\sigma}{\bar{x}} \times 100$$

It is independent of unit, so, two distributions can betterly be compared with the help of C.V. for their variability. Less the C.V., more will be the uniformity, consistency etc. and more the C.V. less will be the uniformity, consistency etc.

ILLUSTRATIVE EXAMPLES

EXAMPLE 1:

Find the range and the coefficient of range from the following data:

Temp(°c):	0-10	10-20	20-30	30-40	40-50
No. of days:	5	12	25	18	2

Solution:

Range = L – S = 50 – 0 = 50 ;

and Coeff. of range $= \frac{L - S}{L + S} = \frac{50 - 0}{50 + 0} = 1.$

EXAMPLE 2:

You are given two variables A and B. Using quartile deviation, state which is more variable.

A	Mid point :	15	20	25	30	35	40	45
	Frequency :	15	33	56	103	40	32	10
B	Mid point :	100	150	200	250	300	350	400
	Frequency :	8	12	13	22	6	8	2

Website :- https://www.arjun00.com.np

Solution:

COMPUTATION OF QUARTILE DEVIATION

Variable A				Variable B			
Mid point	Class interval	f	C.f.	Mid point	Class interval	f	C.f
15	12.5-17.5	15	15	100	75-125	8	8
20	17.5-22.5	33	48	150	125-175	12	20
25	22.5-27.5	56	104	200	175-225	13	33
30	27.5-32.5	103	207	250	225-275	22	55
35	32.5-37.5	40	247	300	275-325	6	61
40	37.5-42.5	32	279	350	325-375	8	69
45	42.5-47.5	10	289	400	375-425	2	71

For A:

$$\frac{N}{4} = \frac{289}{4} = 72.25$$

$\therefore$ Q_1 lies in the class 22.5-27.5

Here, $l = 22.5$, $\frac{N}{4} = 72.25$, c.f. = 48, f = 56, h = 5, $Q_1 = ?$

Now, $$Q_1 = l + \frac{\frac{N}{4} - c.f.}{f} \times h$$

$$= 22.5 + \frac{72.25 - 48}{56} \times 5$$

$$= 22.5 + 2.165 = 24.665$$

Again, $\frac{3N}{4} = 3 \times 72.25 = 216.75$

$\therefore$ Q_3 lies in the class 32.5-37.5

Here, $l = 32.5$, $\frac{3N}{4} = 216.75$, c.f. = 207, f = 40, h = 5, $Q_3 = ?$

Now, $$Q_3 = l + \frac{\frac{3N}{4} - c.f.}{f} \times h$$

Website :- https://www.arjun00.com.np

$$= 32.5 + \frac{216.75 - 207}{40} \times 5$$

$$= 32.5 + 1.219 = 33.719$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{33.719 - 24.665}{33.719 + 24.665} = \frac{9.054}{58.384}$$

$$= 0.155$$

For B:

$$\frac{N}{4} = \frac{71}{4} = 17.75$$

$\therefore$ Q_1 lies in the class 125–175

Here, $l = 125$, $\frac{N}{4} = 17.75$, c.f. = 8, f = 12, h = 50, $Q_1 = ?$

$$\text{Now,} \quad Q_1 = l + \frac{\frac{N}{4} - c.f}{f} \times h$$

$$= 125 + \frac{17.75 - 8}{12} \times 50$$

$$= 125 + \frac{9.75}{12} \times 50$$

$$= 125 + 40.625 = 165.625$$

Again, $\frac{3N}{4} = 3 \times 17.75 = 53.25$

$\therefore$ Q_3 lies in the class 225 - 275

Here, $l = 225$, $\frac{3N}{4} = 53.25$, c.f. = 33, f = 22, h = 50, $Q_3 = ?$

$$\text{Now,} \quad Q_3 = l + \frac{\frac{3N}{4} - c.f}{f} \times h$$

$$= 225 + \frac{53.25 - 33}{22} \times 50$$

$$= 225 + 46.023 = 271.023$$

Website :- https://www.arjun00.com.np

$$\text{Coeff of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{271.023 - 165.625}{271.023 + 165.625}$$

$$= \frac{105.398}{436.648} = .241$$

Since the coeff. of Q.D. of B is greater than that of A, so B is more variable.

EXAMPLE 3:

The following table gives the marks obtained by 50 students in a certain examination Calculate the median and mean deviation from median of the following data.

Marks:	0-10	10-20	20-30	30-40	40-50
No. of students:	5	8	15	16	6

Solution:

x = Mid value, $d = x - M_d$

COMPUTATION OF MEAN DEVIATION

Marks	x	f	c.f.	/d/	f/d/
0-10	5	5	5	23	115
10-20	15	8	13	13	104
20-30	25	15	28	3	45
30-40	35	16	44	7	112
40-50	45	6	50	17	102
		N = 50			$\Sigma f \mid d \mid = 478$

$$\frac{N}{2} = \frac{50}{2} = 25$$

∴ Median lies in the class 20-30

$$M_d = l + \frac{\frac{N}{2} - c.f.}{f} \times h = 20 + \frac{25 - 13}{15} \times 10 = 28$$

$$\text{M.D. from median} = \frac{\Sigma f/d/}{N} = \frac{478}{50} = 9.56 \text{ marks}$$

Website :- https://www.arjun00.com.np

Alternative method

Let a = 25. This value should lie in the class in which median lies.

COMPUTATION OF MEAN DEVIATION

Marks	x	f	/d/ = /x-25/	f/d/
0-10	5	5	20	100
10-20	15	8	10	80
20-30	25	15	0	0
30-40	35	16	10	160
40-50	45	6	20	120
		N = 50		$\Sigma f/d/ = 460$

$\Sigma F_B = 28$, $\Sigma f_A = 22$ $M_d = 28$

$$\text{M.D. from median} = \frac{1}{N}\left[\Sigma f/d/ + (M_a - a)(\Sigma f_B - \Sigma f_A)\right]$$

$$= \frac{1}{50}\left[460 + (28 - 25)(28 - 22)\right]$$

$$= \frac{1}{50} \times 478 = 9.56$$

Note: This alternative method is known as **short cut method.**

This method of finding M.D. is used if the mean or median or mode from which the deviations are to be taken, be in fraction.

EXAMPLE 4:

Find out the mean and standard deviation from the following data:

x:	10	11	12	13	14
f:	3	12	18	12	2

Solution:

COMPUTATION OF MEAN AND S.D.

x	f	d = x – 12	fd	fd^2
10	3	– 2	– 6	12
11	12	– 1	– 12	12
12	18	0	0	0
13	12	1	12	12
14	3	2	6	12
	N = 48		$\Sigma fd = 0$	$\Sigma fd^2 = 48$

Website :- https://www.arjun00.com.np

$$\bar{x} = a + \frac{\Sigma fd}{N} = 12 + \frac{0}{48} = 12$$

and $$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left[\frac{\Sigma fd}{N}\right]^2} = \sqrt{\frac{48}{48} - 0} = 1$$

EXAMPLE 5:

Weight of a group of individuals are given below in kgs. Find out the mean, standard deviation, coefficient of s.d. and the coefficient of variation:

Weight (kg.) Above:	0	10	20	30	40
No of persons:	100	88	55	25	10

Solution:

COMPUTATION OF MEAN AND STANDARD DEVIATION

Wt. (kg)	Mid. Value (x)	f	$d' = \frac{x-25}{10}$	fd'	fd'^2
0–10	5	12	2	–24	48
10–20	15	33	– 1	– 33	33
20–30	25	30	0	0	0
30–40	35	15	1	15	15
40-50	45	10	2	20	40
		N=100		$\Sigma fd' = -22$	$\Sigma fd'^2 = 136$

$$\bar{x} = a + \frac{\Sigma fd'}{N} \times h = 25 + \frac{-22}{100} \times 10 = 22.8 \text{ kgs.}$$

$$\sigma = h \times \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} = 10 \times \sqrt{\frac{136}{100} - \left(\frac{-22}{100}\right)^2}$$

$$= 10 \times \sqrt{1.36 - 0.0484} = 11.45 \text{ kgs.}$$

$$\text{Coeff. of s.d.} = \frac{\sigma}{\bar{x}} = \frac{11.45}{22.8} = 0.502$$

and, $$\text{Coeff. of variation} = \frac{\sigma}{\bar{x}} \times 100 = \frac{11.45}{22.8} \times 100 = 50.2\%$$

Website :- https://www.arjun00.com.np

EXAMPLE 6:

Scores of two golfers for 10 rounds were as follows:

Golfer A: 74, 75, 78, 78, 72, 77, 79, 78, 81, 76

Golfer B: 86, 84, 80, 88, 89, 85, 86, 82, 82, 79

Find which golfer may be considered to be a more consistent player.

Solution:

COMPUTATION OF COEFFICIENT OF VARIATIONS

Golfer A			Golfer B		
x	d=x−77	d^2	x	d=x− 85	d^2
74	-3	9	86	1	1
75	-2	4	84	-1	1
78	1	1	80	-5	25
78	1	1	88	3	9
72	-5	25	89	4	16
77	0	0	85	0	0
79	2	4	86	1	1
78	1	1	82	-3	9
81	4	16	82	-3	9
76	-1	1	79	-6	36
	$\Sigma d = -1$	$\Sigma d^2 = 62$		$\Sigma d = -9$	$\Sigma d^2 = 107$

Golfer A:

$$\overline{x} = a + \frac{\Sigma d}{n} = 77 + \frac{-1}{10} = 76.8$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} = \sqrt{\frac{62}{10} - \left(\frac{-2}{10}\right)^2}$$

$$= \sqrt{6.2 - .04} = 2.48$$

$$\text{C.V. of A} = \frac{\sigma}{\overline{x}} \times 100 = \frac{2.48}{76.8} \times 100 = 3.23\%$$

Golfer B:

$$\overline{x} = a + \frac{\Sigma d}{n} = 85 + \frac{-9}{10} = 84.1$$

Website :- https://www.arjun00.com.np

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} = \sqrt{\frac{107}{10} - \left(\frac{-9}{10}\right)^2}$$

$$= \sqrt{10.7 - .81} = 3.17$$

$$\text{C.V. of B} = \frac{\sigma}{\bar{x}} \times 100 = \frac{3.14}{84.1} \times 100 = 3.73\ \%$$

Since the C.V. of golfer A is less than that of B, so A is the consistent player.

EXAMPLE 7:

An analysis of monthly wages paid to the workers in two firms A and B belonging to the same industry give the following results:

	Firm A	Firm B
No. of workers:	600	700
Average monthly wage:	Rs. 182.00	Rs. 178.50
Variance of distribution of wage:	78	98

(i) Which firm A or B has a larger wage bill ?

(ii) In which firm A or B is there greater variability in individual wages ?

Solution:

Firm A

$n = 600,\ \bar{x} = 182,\ \sigma^2 = 78 \qquad \therefore \sigma = 8.83$

Total wage bill $= \Sigma x = n\bar{x} = 600 \times 182 =$ Rs. 109200

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100 = \frac{8.83}{182} \times 100 = 4.85\%$$

Firm B

$n = 700,\ \bar{x} = 178.50,\ \sigma^2 = 98 \therefore a = 9.9$

Total wage bill $= \Sigma x = n\bar{x} = 700 \times 178.50 =$ Rs. 124950

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100 = \frac{9.9}{178.50} \times 100 = 5.55\%$$

(i) From the above results, form B has a larger wage bill.

(ii) Since the C.V. for firm B is greater than that of A, so the distribution of wages in firm B has greater variability.

Website :- https://www.arjun00.com.np

Exercise 6

1. What do you mean by "Dispersion" ? What purpose does a measure of dispersion serve ? Distinquish between absolute and relative measures of dispersion.

2. Explain briefly what is meant by 'dispersion'. State the requisites of a satisfactory measure of dispersion.

3. What are the various measures of variability ? Discuss the merits and demerits of each of them.

4. What are quartiles ? How are they used in measuring dispersion ? In what respect is the quartile deviation superior of range ?

5. Define mean deviation. How is it different from standard deviation ? Why standard deviation is considered as the best measure of dispersion ? Also,, show that standard deviation is independent of origin but not of scale.

6. What are standard deviation and the coefficient of variation ? What are the main objects of coefficient of variation ?

7. In what way measures of variation supplement measure of central tendency ? Under what situation different measures of dispersion are used ?

8. Define standard deviation. Prove that standard deviation is the least possible root mean square deviation.

9. Find the range and the coefficient of range from the following data:

   (i) 25, 28, 32, 32, 36, 48, 44, 45, 50, 50

   (ii)

Income (Rs):	80-120	120-160	160-200	200-240	240-280
Frequency:	30	25	20	15	10

10. From the following table giving the weights of students calculate the semi interquartile range and the coefficient of quartile deviation.

Wts. (kgs)	Frequency	Wts. (kgs)	Frequency
51	25	57	24
53	21	60	22
56	28	67	18

Website :- https://www.arjun00.com.np

11. Compute the following two series of figures in respect of their dispersion by quartile measurements:

Height (in cms):	158	156	162	161	163	164	165
Wt. (in kgs.):	53	51	57	56	50	65	70

12. Find the mean deviation from mean of the following:

i) 5, 7, 10, 12, 6

ii) 40, 44, 54, 60, 62, 64, 70, 80, 90, 96

13. Calculate the mean deviation from the mean for the following data;

size:	2	4	6	8	10	12	14	16
f:	2	2	4	5	3	2	1	1

14. Calculate mean, median and mean deviations from mean and median and their coefficients from the following data:

Age in year	No of persons	Age of years.	No. of persons
10-20	15	50-60	15
20-30	14	60-70	11
30-40	16	70-80	9
40–50	20		

15. Find the standard deviation of:

i) 30, 40, 35, 22, 25, 48, 45

ii) 114, 115, 123, 120, 110, 130, 119, 118, 116, 115

by direct as well as short cut method.

16. Calculate the standard deviation from the following data:

Size:	06	09	12	15	18
Frequency:	07	12	19	10	3

17. Find out the arithmetic mean and standard deviation from the following data:

Variable:	5-10	10-15	15-20	20-25	25-30	30-35
Frequency:	2	9	29	54	11	5

Website :- https://www.arjun00.com.np

18. Find the mean, s.d. and C.V. from the following data:

Wages(Rs)	No. of Workers	Wages(Rs)	No. of Workers
Upto Rs. 10	12	Upto Rs. 50	157
Upto Rs. 20	30	Upto Rs. 60	200
Upto Rs. 30	65	Upto Rs. 70	222
Upto Rs. 40	107	Upto Rs. 80	230

19. An analysis of monthly wages paid to the workers in two firms A and B belonging to the same industry gives the following result:

	Firm A	Firm B
No of workers:	500	600
Average monthly wages:	Rs. 586	Rs. 575
Variance of distribution of wage:	81	100

(i) Which firm A or B has a larger wage bill ?

(ii) In which firm A or B is there greater variability in individual wages ?

(iii) Calculate the average monthly wage and the variance of wage of all the workers of the firms A and B taken together.

(iv) Which firm is more equitable in regard to wages ?

20. Following are the marks obtained by two students A and B in 10 tests of 100 marks each:

Test:	1	2	3	4	5	6	7	8	9	10
Marks of A:	44	80	76	48	52	72	68	56	60	54
Marks of B:	48	75	54	60	63	69	72	51	57	66

(i) Who is better ?

(ii) Who is intelligent ?

(iii) If the consistency of performance is the criteria for awarding a prize, who should get the prize ?

21. Goals scored by two teams A and B in a football season were as follows:

No of goals in a match:		0	1,	2,	3,	4,
No. of match	A:	27	9	8	5	4
	B:	17	9	6	5	3

By calculating the C.V. in each case find which team may be considered more consistent.

Website : https://www.arjun00.com.np

22. From the data given below state which series is more variable ?

Variable:	10-20	20-30	30-40	40-50	50-60	60-70
Series A:	10	18	32	40	22	18
Series B:	18	22	40	32	29	10

23. Following data represent the lives of two models of refrigerators A and B:

Life (years):		0-2	2-4	4-6	6-8	8-10	10-12
Model	A:	5	16	13	7	5	4
	B:	2	7	17	19	9	1

Find the average life of each model. Which model has greater uniformity ?

24. The following are the monthly salaries in rupees of 30 employees in a firm;

140, 139, 126, 114, 100, 88, 62, 77, 99, 103,
108, 129 144, 148, 134, 63, 69, 148, 132, 118,
142, 116. 123, 104, 95, 80, 85, 106, 123, 133,

The firm gave bonus of Rs. 10, 15, 20, 25, 30 and 35 for individuals in the respective salary groups, exceeding 60 but not exceeding 75, exceeding 75 but not exceeding 90 and so no upto exceeding 135 and not exceeding 150. Find average bonus paid and the standard deviation.

25. The analysis of the results of a budget survey of 150 families gave an average monthly expenditure of Rs. 120 on food items with a standard deviation Rs. 15. After the analysis was completed it was noted that the figure recorded for one household was wrongly taken as Rs. 15 instead of Rs. 105. Determine the correct value of the average expenditure and its standard deviation.

26. Find the mean and the standard deviation of the distribution in which the value of x are 1, 2, 3, ..., n; the frequency of each being unity.

27. Prove that $\sigma^2_{mx+c} = \sigma^2_{mx} = m^2\sigma^2_x$

28. Find the mean deviation from the mean and the standard deviation of the A. P. a, a + d, a + 2d, ..., a + 2nd and prove that the latter is greater than the former.

Website :- https://www.arjun00.com.np

29. Show that if the variable takes the value 0, 1, 2, ...; n with frequencies given by the terms of the binomial series

$$^nC_1\, q^{n-1}\, p,\ ^nC_2\, q^{n-2}\, p^2;\ \ldots\ p^n$$

where $p + q = 1$, then the mean square deviation about origin is $npq + n^2p^2$ and the variance in npq.

30. Show that in a discrete series, if the deviations x from the mean M of a variable X are so small that the third and higher powers of x/M and σ/M can be neglected, the following relations are found to hold approximately.

(i) $H = M\left[1 - \frac{\sigma^2}{M^2}\right]$

(ii) Mean of $\sqrt{x} = \sqrt{M}\left[1 - \frac{\sigma^2}{8M^2}\right]$

Answers

(9) i. 25, 0.33 ii. 200, 0.56 (10) 3.5 kgs, 0.062 (11) Weight is more variable than height (12) i. 2.4. ii. 14.6 (13) 2.8 (14) 42.5 42.5, 15.55, 15.55, 0.37, 0.37. (15) i. 9. 165 ii. 5.25 (16) 3.26 (17) A.M .= 21.05, s.d. = 4.88 (18) Rs. 40.43, Rs. 17.26, 42.69% (19) i. B, ii. B, iii. 580, 121. 36; iv. A (20) i. B, ii B. iii. B (21) B (22) B (23) B (24) Rs. 24.5.. Rs. 8.09 (25) Rs. 120.6, Rs. 12.35 (26) $\frac{1}{2}(n+1)$, $\sqrt{\frac{1}{12}(n^2-1)}$

(28) M.D. = $\frac{n(n+1)/d/}{2n+1}$, s.d. = $/d/\sqrt{\frac{n(n+1)}{3}}$

Website :- https://www.arjun00.com.np

Chapter 7

SKEWNESS AND KURTOSIS

7.1 Introduction and definition

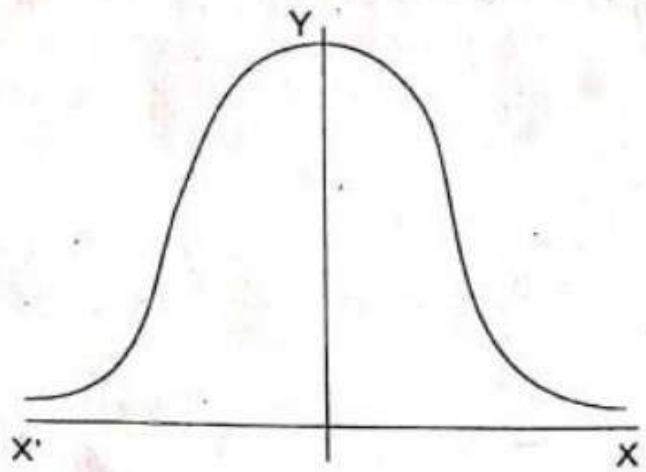
We have discussed about the measures of central tendency and the dispersion. In this chapter, we see whether the dispersed items from the central values are symmetrical or not. A distribution which is not symmetrical is said to be skewed. So, skewness is the lack of symmetry. It gives the degree of departure from symmetry. It relates to the shape of the curve of the frequency distribution than its size. There are two types of skewness; Positive skewness and negative skewness.

Skewness is said to be positive when the curve of the frequency distribution has a longer tail on the right. Also, skewness is positive if mean, median and mode of the frequency distribution satisfy the following condition:

Mean > Median > Mode.

On the other hand, the skewness is negative when the curve has a longer tail on the left or if Mean < Median < Mode.

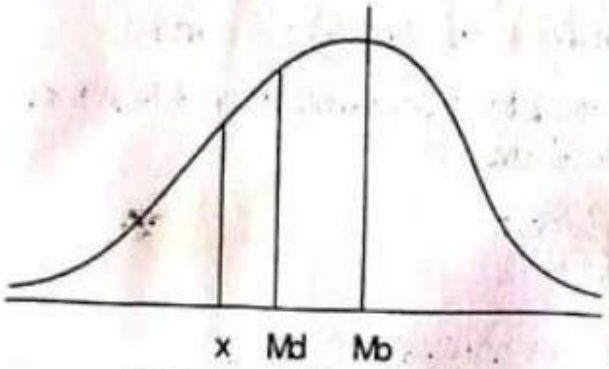
If the frequency distribution is symmetrical, skewness is zero. The following are the figures of symmetrical, positively and negatively skewed distribution.



SYMMETRICAL CURVE

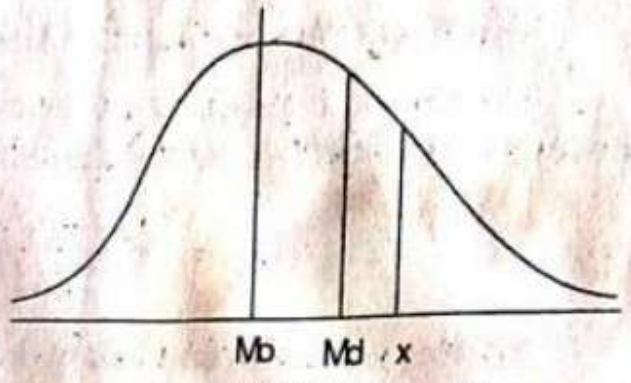
Fig 7.1.

Website :- https://www.arjun00.com.np



NEGATIVELY SKEWED

Fig. 7.2



POSITIVELY SKEWED

Fig. 7.3

The object of studying skewness is to estimate the direction in which and the extend to which the curve of the frequency distribution is away from the symmetrical distribution.

7.2. Measures of Skewness

The absolute measures of skewness are given by the following formulae:

(i) The first measure of skewness

$$= \text{Mean} - \text{Mode or Mean} - \text{Median}$$

(ii) The second measure of skewness

$$= (Q_3 - M_d) - (M_d - Q_1) = Q_3 + Q_1 - 2M_d$$

The first measure of skewness is based on the assumption that in a skew distribution, mean, median and mode do not coincide. But the second measure of skewness depends on the assumption that in a skew distribution, median does not lie in the mid way between lower and upper quartiles. The corresponding relative measures of skewness are given below:

$$\text{First coefficient of skewness} = S_k = \frac{\text{Mean} - \text{Mode}}{\text{s. d.}} = \frac{\overline{x} - M_o}{\sigma}$$

This is also known as Karl Pearson's coefficient of skewness of Pearsonian's coefficient of skewness. As mode is ill defined so it is defficult to locate mode. So, in such a situation, we use the empirical relation (for skew distribution)

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median})$$

$$\text{Then, } S_k = \frac{3(\text{Mean} - \text{Median})}{\text{s. d.}} = \frac{3(\overline{x} - M_d)}{\sigma}$$

Website :- https://www.arjun00.com.np

Karl Pearson's coefficient of skewness lies between –3 and +3.

The second coefficient of skewness based on quartiles is known as Bowley's coefficient of skewness and is defined by

$$S_k = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

Bowley's coefficient of skewness lies between -1 and +1.

From both coefficient of skewness, we can have the following conclusions:

i) If $S_k = 0$, the distribution is symmetrical,

ii) If $S_k > 0$, the distribution is positively skewed.

iii) If $S_k < 0$, the distribution is negatively skewed.

ILLUSTRATIVE EXAMPLES

EXAMPLE 1:

Calculate the mean, standard deviation and the coefficient o skewness from the following frequency distribution.

Investment:	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No of comp:	12	18	20	15	10	3	2

Solution:

COMPUTATION OF COEFFICIENT OF SKEWNESS

Investment	Mid Value	f	$d' = \frac{(x-45)}{10}$	fd'	fd'²
10-20	15	12	-3	-36	108
20-30	25	18	-2	-36	72
30-40	35	20	-1	-20	20
40-50	45	15	0	0	0
50-60	55	10	1	10	10
60-70	65	3	2	6	12
70-80	75	2	3	6	18
		N=80		$\Sigma fd'$ = - 70	$\Sigma fd'^2$ =240

Website :- https://www.arjun00.com.np

For Mean:

$$\overline{x} = a + \frac{\Sigma fd'}{N} \times h = 45 + \frac{-70}{80} \times 10 = 36.25$$

For s.d. :

$$\sigma = h \times \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2}$$

$$= 10 \times \sqrt{\frac{240}{80} - \left(\frac{-70}{80}\right)^2}$$

$$= 10 \times \sqrt{3 - .766} = 14.59 \text{ (Approx.)}$$

By inspection 30-40 is the modal class.

$$\Delta_1 = f_1 - f_0 = 20 - 18 = 2$$

$$\Delta_2 = f_1 - f_2 = 20 - 15 = 5$$

$$M_o = 1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h$$

$$= 30 + \frac{2}{2 + 5} \times 10 = 32.86$$

$$\text{Now, } Sk = \frac{\overline{x} - M_o}{\sigma} = \frac{36.25 - 32.86}{14.95}$$

$$= \frac{3.39}{14.95} = 0.227$$

EXAMPLE 2:

Calculate Karl pearson's coefficient of skewness from the following data:

Marks	Frequency	Marks	Frequency
Above 0	150	Above 50	70
" 10	140	" 60	30
" 20	100	" 70	14
" 30	80	" 80	0
" 40	80		

Website :- https://www.arjun00.com.np

COMPUTATION OF COEFFICIENT OF SKEWNESS

Marks	Mid Value	f	$d' = \frac{(x-45)}{10}$	fd'	fd'^2	c.f.
0-10	5	10	-4	40	160	10
10-20	15	40	-3	-120	360	50
20-30	25	20	-2	-40	80	70
30-40	35	0	-1	0	0	70
40-50	45	10	0	0	0	80
50-60	55	40	1	40	40	120
60-70	65	16	2	32	64	136
70-80	75	14	3	42	126	150
80-90	85	0	4	0	0	150
		N=150		$\Sigma fd' = -86$	$\Sigma fd'^2 = 830$	

$$\bar{x} = a + \frac{\Sigma fd'}{N} \times h = 45 + \frac{-86}{150} \times 10 = 39.27$$

$$\sigma = h \times \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} = 10 \times \sqrt{\frac{830}{150} - \left(\frac{-86}{150}\right)^2}$$

$$= 10 \times \sqrt{5.533 - 0.329} = 22.81$$

Again, $\frac{N}{2} = \frac{150}{2} = 75$, $\therefore$ 40 – 50 is the median class

$$M_d = l + \frac{\frac{N}{2} - c.f.}{f} \times h = 40 + \frac{75 - 70}{10} \times 10 = 45$$

Since the given distribution is the bi-modal distribution so coeff. of skewness

$$S_k = \frac{3(\bar{x} - M_d)}{\sigma}$$

$$= \frac{3(39.27 - 45)}{22.81} = -0.75$$

Website :- https://www.arjun00.com.np

EXAMPLE 3:

Find out the coefficient of skewness based on quartiles for the data given below.

Wages (Rs)	No.of Persons	Wages (Rs)	No. of persons
Less than 36	17	54-59	27
36-41	23	60-65	19
42-47	39	66 and above	12
48-53	35		

Solution:

COMPUTATION OF COEFFICIENT OF SKEWNESS

Wages	No.of persons (f)	c.f.
Less than 36	17	17
36-41	23	40
42-47	39	79
48-53	38	117
54-59	27	144
60-65	19	163
66 and above	12	175

$$\frac{N}{2} = \frac{175}{2} = 87.5, \quad \frac{N}{4} = \frac{175}{4} = 43.75$$

and $$\frac{3N}{4} = \frac{3 \times 175}{4} = 131.25$$

$\therefore$ Median lies in the class 48-53 with real limit 47.5 – 53.5

$$M_d = 1 + \frac{\frac{N}{2} - c.f.}{f} \times h = 47.5 + \frac{87.5 - 79}{15} \times 6 = 48.84$$

Q_1 lies in the class 42-47 with real limit 41.5 – 47.5.

$$Q_1 = 1 + \frac{\frac{N}{4} - c.f.}{f} \times h = 41.5 + \frac{43.75 - 40}{39} \times 6 = 42.08$$

Website :- https://www.arjun00.com.np

Q_3 lies in the class 54 - 59 with real limit 53.5 - 59.5

$$Q_3 = l + \frac{\frac{3N}{4} - c.f.}{f} \times h = 53.5 + \frac{131.25 - 117}{27} \times 6 = 56.67$$

Now, Coefficient of skewness based on quartiles (Bowley's coeff of skewness)

$$= \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} = \frac{56.67 + 42.08 - 2 \times 48.84}{56.67 - 42.08}$$

$$= \frac{98.75 - 97.68}{14.59} = \frac{1.07}{14.59} = 0.073$$

EXAMPLE 4:

In a certain distribution, the following results were obtained Mean = 45, Median = 48, Coefficient of skewness = -0.4 Find s.d., and the coefficient of variation .

Solution:

Here, $\overline{x} = 45$, $M_d = 48$, $S_k = -0.4$, $\sigma = ?$

Now, $S_k = \frac{3(\overline{x} - M_d)}{\sigma}$ or, $-0.4 = \frac{3(45 - 48)}{\sigma}$

or, $-0.4 = \frac{-9}{\sigma}$ $\therefore \sigma = 22.5$

Again C.V. $= \frac{\sigma}{\overline{x}} \times 100 = \frac{22.5}{45} \times 100 = 50\%$

EXAMPLE 5:

The following facts were gathered before and after an industrial dispute:

	Before dispute	After dispute
No of workers employed:	516	508
Mean wage:	49.50	51.75
Median wage:	52.70	50.00
Variance of wage:	100	121

Compare the position before and after the dispute in respect of a) total wages b) modal wages c) coefficient of variation d) skewness.

Website :- https://www.arjun00.com.np

Solution:

The comparisons can be made under the basis of the information given.

a) Total wage bill before dispute = $n\overline{x} = 516 \times 49.50$

$= \text{Rs. } 25542$

" " after dispute = $n\overline{x} = 508 \times 51.75$

$= \text{Rs. } 26289$

After the dispute,the total wage bill has been increased by Rs. 747, though the numbers of workers has decreased by 8. This is due to the fact that mean wage after the dispute has increased which will certainly be beneficial to the workers. The increase in the wage bill should not be considered as a disadvantage to the management if there is an increase in the efficiency of the workers and increase in productivity.

b) Before dispute mode = 3Median - 2Mean

$= 3 \times 52.7 - 2 \times 49.50 = 59.1$

After dispute mode = 3Median - 2Mean

$= 3 \times 50 - 2 \times 51.75 = 46.5$

Thus we see that the modal wage has come down from Rs. 59.1 to Rs. 46.50 So, after dispute, there is a concentration of wages around a small value.

C.V. before dispute $= \frac{\sigma}{\overline{x}} \times 100 = \frac{10}{49.50} \times 100 = 20.2\%$

C.V. after dispute $= \frac{\sigma}{\overline{x}} \times 100 = \frac{11}{51.75} \times 100 = 21.26\%$

Since the C. V. after dispute is higher than before, so wage distribution after dispute is not uniform.

d) Coeff. of skewness before dispute $= \frac{3(\overline{x} - M_d)}{\sigma}$

$= \frac{3(59.50 - 52.70)}{10} = -0.96$

Coeff. of skewness after dispute $= \frac{3(\overline{x} - M_d)}{\sigma}$

$= \frac{3(51.75 - 50)}{11} = 0.48.$

Website :- https://www.arjun00.com.np

8 Unit

Bivariate Data Analysis

8.1 Introduction

So far we have confined our discussion to the distribution involving only one variable (univariate distribution) and also saw how the various measures of central tendency, dispersion and shape of frequency distribution or curve can be used for the purpose of comparison and analysis. We may, however, come across certain series where each item of the series may assume the values of two or more variables.

Sometimes two continuous characters are measured in the same person such as weight and cholesterol, weight and height, age and weight, age and height etc. At other times, the same character is measured in two related groups such as tallness in parents and tallness in children, study of intelligent quotient (IQ) in brothers and in corresponding sisters and so on. The relationship between two quantitatively measured variables are called correlation and regression. ***Thus the correlation and regression are statistical tools which study the relationship between two or more than two variables.***

8.2 Correlation and Its Types

Correlation is described in different ways. Three of most important ways of classifying correlation are:

1. Positive and negative correlation
2. Linear and non-linear correlation
3. Simple, partial and multiple correlation

1. Positive and negative correlation

The correlation between two variables is said to be positive, if the increase (or decrease) in the value of one variable is accompanied by an increase (or decrease) in the value of the other variable. For example, consider the data

X:	5	9	17	20	32
Y:	3	6	8	15	16

Website :- https://www.arjun00.com.np

Here the value of Y has increased corresponding to every increasing value of X. The correlation between the variables X and Y is positive. The following data also reveals the positive correlation between two variables X and Y.

X:	27	16	13	9	5
Y:	9	7	4	3	1

The correlation between two variables is said to be negative, if increase (or decrease) in the value of one variable is accompanied by a decrease (or increase) in the value of the other variable. For example, consider the data

X:	115	107	104	98	70
Y:	25	27	43	49	64

Here a decrease in the value of X is accompanied by an increase in the value of Y. The correlation between X and Y is negative.

2. Linear and non-linear correlation

The correlation between two variables is said to be linear, if the ratio of change in one variable to the change in the other variable is almost constant. For example, consider the data,

X:	1	2	3	4	5
Y:	9	11	13	15	17

Here the variables X and Y is linear.

The correlation between two variables is said to be non-linear, if the ratio of change in one variable to the change in the other variable is not constant. Let us consider the following data:

X:	1	2	3	4	5
Y:	7	10	23	30	42

Here the correlation between the variables X and Y is non-linear.

3. Simple, multiple and partial correlation

The correlation between two variables is known as simple correlation, when three or more variables are considered, then the correlation may be multiple or partial. The degree of the relationship among three or more variables at a time is called multiple correlation. In other words, in multiple correlation all the given variables are studied at a time by keeping one variable as dependent and all the remaining variables as independent. Thus in multiple correlation three or more variables studied simultaneously. For example, when we study the relationship between the yield of rice per hector and both the amount of rainfall and the amount of fertilizers used, it is a problem of multiple correlation.

In partial correlation, there will be three or more variables. The degree of relationship between dependent variable and one independent variable by keeping all other independent variables as constant is called

Website :- https://www.arjun00.com.np

partial correlation. For example, the achievement of the child in the school is affected by the age, sex, IQ, number of hours studied at home, quality of teaching etc. When we study the relationship between the achievement of the child and age by keeping all other variable as constant is a problem of partial correlations.

8.3 Methods of Measuring Correlation

The following are the methods of measuring correlation between two variables.

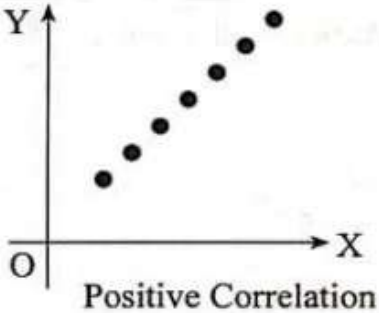
1. Scatter diagram method
2. Karl Pearson's method
3. Spearman's rank correlation

Scatter diagram method:

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be n pairs of values of two variables X and Y with respect to some characteristic. If we plot the corresponding values of X and Y, taking the values of X along x-axis and the values of Y along y-axis, we obtain a collection of dots, called dot diagram or a scatter diagram.

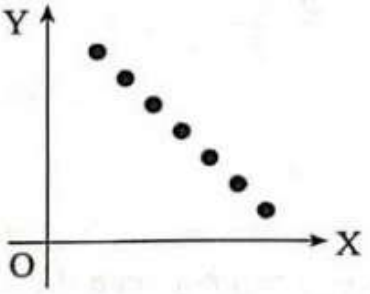
The correlation between two variables by scatter diagram can be analyzed in the following manner:

(a) If the plotted points are in a straight line raising from lower left to upper right, then there is positive correlation between the variables.



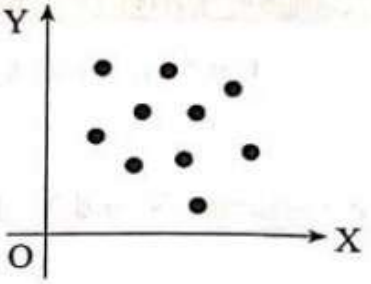
Positive Correlation

(b) If the plotted points are in a straight line falling from upper left to lower right, then there is negative correlation between the variables.



Negative Correlation

(c) If the plotted points show no trend at all, then the variables are independent and are not correlated.



Website :- https://www.arjun00.com.np

Merits and demerits of scatter diagram method

Merits

1. This is the simplest method.
2. This method is most suitable, if just an idea is to be formed whether the variables are correlated or not.
3. This method is least affected by the extreme items.
4. Scatter diagrams are attractive and can be used for in exhibition displaying the correlation between variables.

Demerits

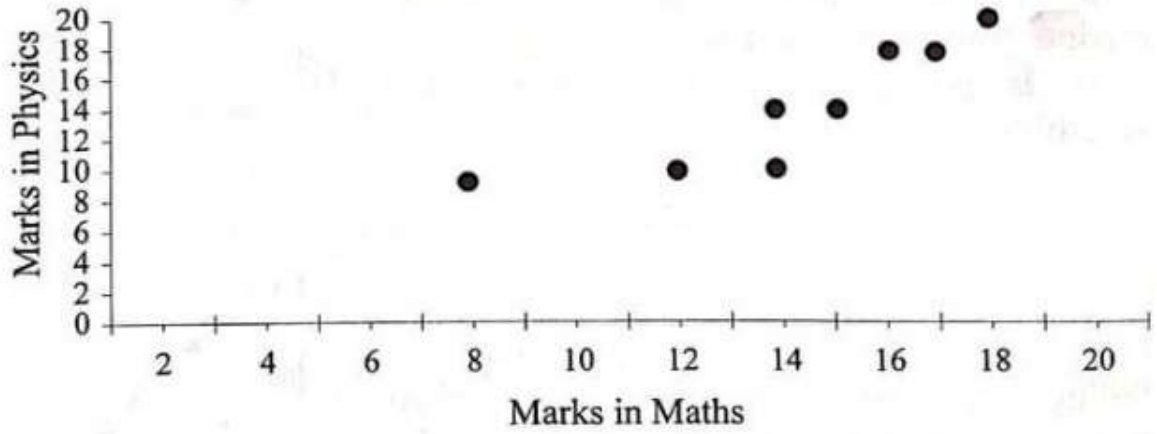
1. This method is subjective in nature.
2. The degree of correlation assessed, cannot be numerically expressed.

Example 1.

Draw a scatter diagram for the following data and comment on the presence of correlation between the variables:

Marks in maths	8	12	16	15	14	17	18
Marks in physics	9	10	18	14	14	18	20

Scatter Diagram



From the scatter diagram, we see that there is positive linear correlation between the marks obtain in mathematics and physics.

8.4 Covariance

Let $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots, (X_n, Y_n)$ be n pairs of values of two variables X and Y. The measure $\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n}$ is defined as the

Website :- https://www.arjun00.com.np

covariance between the variables X and Y. This is denoted by cov(X, Y). For simplicity, we write

$$\text{Cov}(X, Y) = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n}$$ This formula is used when $\overline{X}$ and $\overline{Y}$ are in whole number.

Let us derive another expression for covariance.

We know that

$$\text{Cov}(X, Y) = \frac{\Sigma(x_i-\overline{x})(y_i-\overline{y})}{n} = \frac{1}{n}\{\Sigma(x_i-\overline{x})(y_i-\overline{y})\}$$

$$= \frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+ \ldots\ldots + (x_n-\overline{x})(y_n-\overline{y})\}$$

$$= \frac{1}{n}\{x_1y_1 - x_1\overline{y} - \overline{x}y_1 + \overline{x}\,\overline{y} + x_2y_2 - x_2\overline{y} - \overline{x}y_2 + \overline{x}\,\overline{y} + \ldots + x_ny_n - \overline{x}y_n - x_n\overline{y} + \overline{x}\,\overline{y}\}$$

$$= \frac{1}{n}\{(x_1y_1 + x_2y_2 + \ldots + x_ny_n) - (x_1 + x_2 + \ldots + x_n)\overline{y} - (y_1 + y_2 + \ldots + y_n)\overline{x} + (\overline{x}\,\overline{y} + \overline{x}\,\overline{y} + \ldots + \overline{x}\,\overline{y})\}$$

$$= \frac{1}{n}\left\{\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_i\overline{y} - \sum_{i=1}^{n}y_i\overline{x} + n\overline{x}\,\overline{y}\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}\frac{x_i}{n}\overline{y} - \sum_{i=1}^{n}\frac{y_i}{n}\overline{x} + \overline{x}\,\overline{y}$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{x}\,\overline{y} - \overline{x}\,\overline{y} + \overline{x}\,\overline{y} \quad \left[\because \sum_{i=1}^{n}\frac{x_i}{n} = \overline{x} \;\&\; \sum_{i=1}^{n}\frac{y_i}{n} = \overline{y}\right]$$

$$\therefore \text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{x}\,\overline{y}$$

This formula is used when the $\overline{x}$ and $\overline{y}$ are not whole number (i.e. if $\overline{x}$, $\overline{y}$ are in fraction or decimals).

Example 2.

Find Cov(X, Y) from the following data:

X	15	20	25	30	40	50
Y	44	43	45	37	34	37

Website :- https://www.arjun00.com.np

Solution:

CALCULATION OF COVARIANCE

X	Y	$X-\overline{X}$	$Y-\overline{Y}$	$(X-\overline{X})(Y-\overline{Y})$
15	44	– 15	4	– 60
20	43	– 10	3	– 30
25	45	– 5	5	– 25
30	37	0	– 3	0
40	34	10	– 6	– 60
50	37	20	– 3	– 60
$\Sigma X = 180$	$\Sigma Y = 240$			$\Sigma(X-\overline{X})(Y-\overline{Y}) = -235$

Now,

$$\overline{X} = \frac{\Sigma x}{N} = \frac{180}{6} = 30$$

$$\overline{Y} = \frac{\Sigma y}{N} = \frac{240}{6} = 40$$

$$\therefore \text{Cov}(X, Y) = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n} = \frac{-235}{6} = -39.167.$$

Example 3.

Find the covariance between x and y for the following observations of (x, y):

(2, 4), (10, 15), (20, 12), (10, 16), (17, 17), (12, 10), (15, 11), (16, 8).

Solution:

CALCULATION OF COVARIANCE

X	Y	XY
2	4	8
10	15	150
20	12	240
10	16	160
17	17	289
12	10	120
15	11	165
16	8	128
$\Sigma X = 102$	$\Sigma Y = 93$	$\Sigma XY = 1260$

Website :- https://www.arjun00.com.np

We know that

$$\overline{X} = \frac{\Sigma X}{n} = \frac{102}{8} = 12.75$$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{93}{8} = 11.63$$

Since $\overline{x}$ and $\overline{y}$ are not whole number so,

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \overline{X}\,\overline{Y} = \frac{1}{8}(1260) - (12.75)(11.63)$$

$$= 9.29$$

8.5 Karl Pearson's Coefficient of Correlation

A mathematical method for measuring the degree of linear relationship between two variables was given by Karl Pearson (1867 – 1936), a great British Bio-matrician and statistician. Karl Pearson's measure of correlation coefficient between two variables X and Y, usually denoted by r_{XY} or simply r, is a numerical measure of linear relationship between them and is defined as the ratio of the cov(X, Y) to the product of square roots of variances of X and Y.

$$\therefore r_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}X}\sqrt{\text{Var}Y}} \qquad (1)$$

$$\text{Since Cov}(X, Y) = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n}$$

$$\text{Var}(X) = \frac{\Sigma(X-\overline{X})^2}{n}$$

$$\text{and Var}(Y) = \frac{\Sigma(Y-\overline{Y})^2}{n}$$

From (1), we write

$$r_{XY} = \frac{\dfrac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n}}{\sqrt{\dfrac{\Sigma(X-\overline{X})^2}{n}}\sqrt{\dfrac{\Sigma(Y-\overline{Y})^2}{n}}} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{\Sigma(Y-\overline{Y})^2}}$$

(2)

For simplicity, relation (2) can be written as

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} \qquad (3)$$

Website :- https://www.arjun00.com.np

where $x = X - \overline{X}$ and $y = Y - \overline{Y}$.

From (2) after simplification we can write

$$r_{XY} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} \qquad (4)$$

This is known as **direct method**. The formula (4) facilitates to compute the coefficient of correlation of the ungrouped data of small numbers directly from the original data.

Example 4.

Find the coefficient of correlation between x and y, when $Cov(X, Y) = 16.5$, $Var(X) = 2.89$ and $Var(Y) = 100$.

Solution:

$$\text{We have, } r = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{16.5}{\sqrt{2.89}\sqrt{100}} = \frac{16.5}{\sqrt{289}}$$

$$= \frac{16.5}{17} \qquad \therefore \quad r = 0.97$$

Example 5.

Find out the coefficient of correlation between the heights of the fathers and sons:

Height of fathers(in inches):	65	66	67	68	69	70	71
Height of sons(in inches):	67	68	66	69	72	72	69

Solution:

COMPUTATION OF CORRELATION COEFFICIENT

Height of fathers X	Height of sons Y	$x = X - \overline{X}$	$y = Y - \overline{Y}$	x^2	y^2	xy
65	67	−3	−2	9	4	6
66	68	−2	−1	4	1	2
67	66	−1	−3	1	9	3
68	69	0	0	0	0	0
69	72	1	3	1	9	3
70	72	2	3	4	9	6
71	69	3	0	9	0	0
$\Sigma X = 476$	$\Sigma Y = 483$			$\Sigma x^2 = 28$	$\Sigma y^2 = 32$	$\Sigma xy = 20$

Website :- https://www.arjun00.com.np

Now,

Average height of fathers

$$\overline{X} = \frac{\Sigma X}{N} = \frac{476}{7} = 68$$

Average height of sons

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{483}{7} = 69$$

We know,

$$\text{Correlation coefficient } (r_{XY}) = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$$

$$= \frac{20}{\sqrt{28}\sqrt{32}} = \frac{20}{\sqrt{896}} = \frac{20}{29.93}$$

$$= 0.67$$

Therefore, there is positive correlation between the heights of the fathers and sons.

Example 6.

Calculate the coefficient of correlation between X and Y series from the following data:

	Series X	Series Y
No. of observations	20	20
s.d.	3.25	3.5

$$\Sigma(X - \overline{X})(Y - \overline{Y}) = 120$$

Solution:

We have,

$$r = \frac{Cov(X, Y)}{\sqrt{Var X}\sqrt{Var Y}} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n\,\sigma_X\,\sigma_Y}$$

$$= \frac{120}{20(3.25)(3.5)} = \frac{120}{227.5} = 0.53$$

Example 7.

Calculate the coefficient of correlation from the following data:

X	5	2	10	4	7	8	17	6
Y	9	13	5	7	11	7	7	12

Website :- https://www.arjun00.com.np

Solution:

COMPUTATION OF CORRELATION COEFFICIENT

X	Y	X^2	Y^2	XY
5	9	25	81	45
2	13	4	169	26
10	5	100	25	50
4	7	16	49	28
7	11	49	121	77
8	7	64	49	56
17	7	289	49	119
6	12	36	144	72
$\Sigma X = 59$	$\Sigma Y = 71$	$\Sigma X^2 = 583$	$\Sigma Y^2 = 687$	$\Sigma XY = 473$

$$\text{Now, } r_{XY} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{8 \times 473 - 59 \times 71}{\sqrt{8 \times 583 - (59)^2}\sqrt{8 \times 687 - (71)^2}}$$

$$= -0.55$$

Step deviation method

If $\overline{X}$ and $\overline{Y}$ are fractional and X and Y assume large values, the above formula are not generally used for numerical problems. In such a case, the step deviation method where we take the deviations of the variables X and Y from any arbitrary numbers, is used. For this we use the following formula,

$$r = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

The values of u and v are the step deviations of the values of X and Y respectively. To calculate r_{XY}, the following steps should be adopted

(a) Write the values of X and Y. Find deviation X – A and Y – B from suitable constants A and B respectively. If there is some common factor h(>1) in the value of X – A then find $\frac{X-A}{h}$ in

Website :- https://www.arjun00.com.np

the next column and put $\frac{X-A}{h} = u$. If there is no common factor other them 1 in the value of X – A, then put this itself as u. In the similar way, find the value of the variable v as $v = \frac{Y-B}{k}$.

(b) Make the columns with headings uv, u^2 and v^2.

(c) Find the sum of values of u, v, uv, u^2 and v^2.

(d) Calculate $r = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$

Example 8.

Calculate the coefficient of correlation from the following data:

X	14	16	19	22	24	30
Y	24	22	20	24	23	26

Solution:

COMPUTATIONS OF CORRELATION COEFFICIENT

X	Y	$u = X - 19$	$v = Y - 23$	u^2	v^2	uv
14	24	–5	1	25	1	–5
16	22	–3	–1	9	1	3
19	20	0	–3	0	9	0
22	24	3	1	9	1	3
24	23	5	0	25	0	0
30	26	11	3	121	9	33
		$\Sigma u = 11$	$\Sigma v = 1$	$\Sigma u^2 = 189$	$\Sigma v^2 = 21$	$\Sigma uv = 34$

We know that,

$$r = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}} = \frac{6\times 34 - 11\times 1}{\sqrt{6\times 189 - (11)^2}\sqrt{6\times 21 - (1)^2}}$$

$$= \frac{204-11}{\sqrt{1134-121}\sqrt{126-1}} = \frac{193}{\sqrt{1013}\sqrt{125}} = \frac{193}{\sqrt{126625}} = \frac{193}{355.84}$$

$\therefore r = 0.542$

Website :- https://www.arjun00.com.np

8.6 Correlation Coefficient of Bi-variate Distribution

If values of two variables (or series) X and Y are grouped with their corresponding given frequencies of different groups which are presented combinaly in a table is known as "Bivariate frequency table" or "Correlation table". This table is constructed if the number of observations is too large and the data have been classified according as two measurements. The correlation coefficient of bi-variate distribution is obtained by using the formula as:

$$r = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{\sqrt{N\Sigma fu^2 - (\Sigma fu)^2}\sqrt{N\Sigma fv^2 - (\Sigma fv)^2}} \text{ where,}$$

$$N = \Sigma f,\ u = \frac{x-a}{h} \text{ and } v = \frac{y-b}{k}.$$

Steps to obtain the correlation coefficient of bi-variate distribution

1. Obtain middle values of each class group of X and Y and denote it respectively by x and y.
2. Obtain $u = \frac{x-a}{h}$ and $v = \frac{y-b}{k}$ where a, b are two assumed means of mid-values x, y and h, k be the class intervals of two series X, Y respectively.
3. Find total frequency of all classes of y goes horizontally, kept in lower left corner of each small rectangular cell for each interval of X and again similarly find total frequency of all class intervals of X goes vertically, kept in upper right corner of each small rectangular cell.
4. Find fv, fv^2 and kept horizontally below each class interval of Y and find fu, fu^2 and kept vertically, right to each class interval of X again find their corresponding sums also.
5. Find fuv for each class interval of Y taking vertical sum of the values kept at the upper right corner of each cell and again find fuv for each class interval of X taking horizontal sum of the values kept at the upper right corner of each cell also find their corresponding sums horizontally and vertically respectively.
6. Obtain the correlation coefficient (r) of bi-variate distribution by using the above formula.

Website :- https://www.arjun00.com.np

Example 9.

From the data of marks obtained by students in Math and Physics, find the Karl Pearson correlation coefficient and interpret the result.

Mark in Math	Marks in Physics				
	20-30	30-40	40-50	50-60	60-70
25-35	5	9	3	–	–
35-45	2	10	25	2	–
45-55	1	2	12	2	3
55-65	–	–	4	16	5
65-75	–	–	–	4	2

[CTEVT 2073 R/B]

Solution:

Computation of Correlation Coefficient

		y ⇒	Marks in Physics	20-30	30-40	40-50	50-60	60-70				
x⇓			Mid-value of Y	25	35	45	55	65				
Marks in Maths	Mid-value of X	u	v	–2	–1	0	1	2	f	fu	fu^2	fuv
25-35	30	–2		20 5	18 9	0 3	–	–	17	–34	68	38
35-45	40	–1		4 2	10 10	0 25	–2 2	–	39	–39	39	12
45-55	50	0		0 1	0 2	0 12	0 2	0 3	20	0	0	0
55-65	60	1		–	–	0 4	16 16	10 5	25	25	25	26
65-75	70	2		–	–	–	8 4	8 2	6	12	24	16
			f	8	21	44	24	10	107	–36	156	92
			fv	–16	–21	0	24	20	7			
			fv^2	32	21	0	24	40	117			
			fuv	24	28	0	22	18	92			

$$r = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{\sqrt{N\Sigma fu^2 - (\Sigma fu)^2}\sqrt{N\Sigma fv^2 - (\Sigma fv)^2}}$$

Website :- https://www.arjun00.com.np

$$= \frac{107\times 92-(-36)(7)}{\sqrt{107\times 156-(-36)^2}\sqrt{107\times 117-(7)^2}}$$

$$= \frac{9844+252}{\sqrt{16692-1296}\sqrt{12519-49}}$$

$$= \frac{10096}{\sqrt{15396}\sqrt{12470}}$$

$$= \frac{10672}{124.081\times 111.67} = \frac{10672}{13856.020}$$

$$= 0.770$$

$\therefore$ The interpretation of result is that there is high degree positive correlation between the marks of Math and Physics.

8.7 Properties of Correlation Coefficient

Some important properties of the correlation coefficients and their proof are as follows:

1. The correlation coefficient between two variables lies in between – 1 and + 1. i.e. $-1 \le r \le 1$

Proof:

Let us consider the sum of squares

$$\sum\left[\frac{x-\bar{x}}{\sigma_x} \pm \frac{y-\bar{y}}{\sigma_y}\right]^2$$, which is always non-negative.

i.e. $$\sum\left[\frac{x-\bar{x}}{\sigma_x} \pm \frac{y-\bar{y}}{\sigma_y}\right]^2 \ge 0$$

$$\Rightarrow \quad \sum\left[\left(\frac{x-\bar{x}}{\sigma_x}\right)^2 + \left(\frac{y-\bar{y}}{\sigma_y}\right)^2 \pm 2\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right)\right] \ge 0$$

$$\Rightarrow \quad \frac{\Sigma(x-\bar{x})^2}{\sigma_x^2} + \frac{\Sigma(y-\bar{y})^2}{\sigma_y^2} \pm \frac{2\Sigma(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} \ge 0$$

Dividing throughout by n, the number of pairs of observations, we get

Website :- https://www.arjun00.com.np

$$\frac{1}{\sigma_x^2}.\frac{1}{n}\Sigma(x-\bar{x})^2 + \frac{1}{\sigma_y^2}.\frac{1}{n}\Sigma(y-\bar{y})^2 \pm$$

$$\frac{2}{\sigma_x\sigma_y}.\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y}) \geq 0$$

$$\Rightarrow \quad \frac{1}{\sigma_x^2}.\sigma_x^2 + \frac{1}{\sigma_y^2}.\sigma_y^2 \pm \frac{2}{\sigma_x\sigma_y}\text{Cov}(x, y) \geq 0$$

$$\text{But } r = \frac{\text{Cov}(x, y)}{\sigma_x\sigma_y}$$

$$\Rightarrow \quad 1 + 1 \pm 2r \geq 0$$

$$\Rightarrow \quad 2 \pm 2r \geq 0$$

$$\Rightarrow \quad 1 \pm r \geq 0$$

Taking +ve sign

$$1 + r \geq 0$$

$$\Rightarrow \quad -1 \leq r \qquad (1)$$

Taking –ve sign

$$1 - r \geq 0$$

$$\Rightarrow \quad r \leq 1 \qquad (2)$$

Combining (1) and (2), we get

$$-1 \leq r \leq 1$$

2. The correlation coefficient between two variables is independent of change of origin and scale.

Proof:

Let X and Y be two given variables and r_{XY} be the correlation coefficient between them.

$$\text{Then, } r_{XY} = \frac{\Sigma(X-\bar{X})\ (Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{\Sigma(Y-\bar{Y})^2}} \qquad (1)$$

Let the original variables X and Y be transformed to the new variables U and V by the change of origin and scale.

$$\text{i.e., } U = \frac{X-A}{h} \text{ and } V = \frac{Y-B}{k} \qquad (2)$$

where A, B, h and k are constant and $h > 0$, $k > 0$. We have to show that the correlation coefficient between X and Y is same as the correlation coefficient between U and V.

Website :- https://www.arjun00.com.np

i.e., $r_{XY} = r_{UV}$

From (2),

$$X = A + hU \text{ and } Y = B + kV \tag{3}$$

Summing both sides and dividing by n, we get

$$\overline{X} = A + h\overline{U} \quad \text{and } \overline{Y} = B + k\overline{V} \tag{4}$$

Subtracting (4) from (3), we get

$$X - \overline{X} = h(U - \overline{U}) \text{ and } Y - \overline{Y} = k(V - \overline{V})$$

Substituting these values in (1), we get

$$r_{XY} = \frac{\Sigma h(U-\overline{U}) \times k(V-\overline{V})}{\sqrt{\Sigma h^2 (U-\overline{U})^2}\sqrt{\Sigma k^2 (V-\overline{V})^2}}$$

$$= \frac{hk\Sigma(U-\overline{U})\ (V-\overline{V})}{hk\sqrt{\Sigma(U-\overline{U})^2}\sqrt{(\Sigma(V-\overline{V})^2}}$$

$$= \frac{\Sigma(U-\overline{U})\ (V-\overline{V})}{\sqrt{\Sigma(U-\overline{U})^2}\sqrt{\Sigma(V-\overline{V})^2}}$$

$= r_{UV}$. Hence proved.

3. The correlation coefficient between the two variables X and Y is symmetrical. i.e. $r_{XY} = r_{YX}$.

8.8 Interpretation of Correlation Coefficient

The value of Karl person's coefficient of correlation can be interpreted as follows:

1. If $r = +1$ then there is perfect positive linear correlation.
2. If $0.75 \le r < 1$ then there is high degree positive linear correlation.
3. If $0.50 \le r < 0.75$ then there is the moderate degree positive linear correlation.
4. If $0 < r < 0.50$ then there is low degree positive linear correlation.
5. If $r = 0$ then there is no linear correlation.

For the negative value of r can be interpreted as negative correlation as mentioned above.

Website :- https://www.arjun00.com.np

EXAMPLE 4:

The following table gives bivariate frequency distribution of 50 clerks according to age in years and pay in rupees.

Age	Pay			
	250-300	300-350	350-400	400-450
20-30	8	3	–	–
30-40	2	5	2	2
40-50	–	2	9	6
50-60	-	-	5	6

Website :- https://www.arjun00.com.np

Solution:

Let the age and the pay be denoted by X and Y respectively.

COMPUTATION OF CORRELATION COEFFICIENT

$d_x = \frac{x-45}{10}$, $d_y = \frac{y-375}{50}$

x \ y		Pay	250-300	300-350	350-400	400-450				
		y Mid-point	275	325	375	425				
Age	x Mid-point	d_x \ d_y	−2	−1	0	1	f	fu	fu^2	fuv
20–30	25	−2	$\boxed{32}$ 8	$\boxed{6}$ 3			11	−22	44	38
30–40	35	−1	$\boxed{4}$ 2	$\boxed{5}$ 5	$\boxed{0}$ 2	$\boxed{-2}$ 2	11	−11	11	7
40–50	45	0		$\boxed{0}$ 2	$\boxed{0}$ 9	$\boxed{0}$ 6	17	0	0	0
50–60	55	1			$\boxed{0}$ 5	$\boxed{6}$ 6	11	11	11	6
		f	10	10	16	14	N = 50	Σfu = −22	Σfu^2 = 66	Σfuv = 51
		fv	−20	−10	0	14	Σfv = −16			
		fv^2	40	10	0	14	Σfv^2 = 64			
		fuv	36	11	0	4	Σfuv = 51			

Here, $N = 50$, $\Sigma fu = -22$, $\Sigma fu^2 = 66$, $\Sigma fv = -16$ $\Sigma fv^2 = 64$, $\Sigma fuv = 51$, $r = ?$

Now,
$$r = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{\sqrt{N\Sigma fu^2 - (\Sigma fu)^2}\ \sqrt{N\Sigma fv^2 - (\Sigma fu)^2}}$$
$$= \frac{50 \times 51 - (-22) \times (-16)}{\sqrt{50 \times 66 - (-22)^2}\ \sqrt{50 \times 64 - (16)^2}}$$
$$= \frac{2550 - 352}{\sqrt{2816 \times 2950}} = \frac{2198}{28882.22} = 0.763$$

There is a high degree positive correlation between the age and the pay.

Website :- https://www.arjun00.com.np

EXAMPLE 6:

A computer while calculating correlation coefficient between two variables x and y from 25 pair of observations obtained the following results.

$n = 25, \Sigma x = 125, \Sigma x^2 = 650, \Sigma y = 100, \Sigma y^2 = 460, \Sigma xy = 508$. It was however,later discovered at the time of checking that he had copied down pairs as $\begin{array}{c|c} x & y \\ \hline 6 & 14 \\ 8 & 6 \end{array}$ while the correct values were $\begin{array}{c|c} x & y \\ \hline 8 & 12 \\ 6 & 8 \end{array}$. Obtain the correct value of the correlation coefficient.

Solution:

$$n = 25$$

Corrected $\Sigma x = 125 - 6 - 8 + 8 + 6 = 125$

" $\Sigma y = 100 - 14 - 6 + 12 + 8 = 100$ corrected

" $\Sigma x^2 = 650 - (6)^2 - (8)^2 + (8)^2 + (6)^2 = 650$ "

" $\Sigma y^2 = 460 - (14)^2 - (6)^2 + (12)^2 + (8)^2 = 436$"

" $\Sigma xy = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8$

$= 520$

r_c = Corrected value or r = ?

Now,

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

$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2}\sqrt{25 \times 436 - (100)^2}}$$

$$= \frac{500}{(625 \times 900)^{1/2}} = \frac{500}{25 \times 30} = \frac{2}{3}$$

Website :- https://www.arjun00.com.np

EXAMPLE 7 :

The variates x and y have zero means, the same variance σ^2 and zero correlation. Show that the variables

$$u = x\cos\alpha + y\sin\alpha \qquad \text{and} \qquad v = x\sin\alpha - y\cos\alpha$$

have the same variance σ^2 and zero correlation.

Solution:

$$\overline{x} = 0, \qquad \overline{y} = 0, \qquad \Sigma(x-\overline{x})(y-\overline{y}) = \Sigma xy = 0$$

Also, $\overline{u} = 0$ and $\overline{v} = 0$, $\sigma_x^2 = \sigma_y^2 = \sigma^2$

Now, $\sigma_u^2 = \frac{1}{n}\Sigma u^2 = \frac{1}{n}\Sigma(x\cos\alpha + y\sin\alpha)^2$

$$= \cos^2\alpha.\frac{1}{n}\Sigma x^2 + \sin^2\alpha\,\frac{1}{n}\Sigma y^2 + 0$$

$$= \cos^2\alpha.\sigma_x^2 + \sin^2\alpha\,\sigma_y^2$$

$$= \sigma^2(\cos^2\alpha + \sin^2\alpha) = \sigma^2$$

Similarly, $\sigma_v^2 = \sigma^2$

Again $\Sigma uv = \Sigma(x\cos\alpha + y\sin\alpha)(x\sin\alpha - y\cos\alpha)$

$$= \sin\alpha\cos\alpha\Sigma x^2 - \cos^2\alpha\,\Sigma xy + \sin^2 x\,\Sigma xy - \sin\alpha\cos\alpha\,\Sigma y^2$$

$$= \sin\alpha\cos\alpha\, n\sigma_x^2 - 0 + 0 - \sin\alpha\cos\alpha\, n\sigma y^2$$

$$= 0 \qquad (\therefore \sigma_x^2 = \sigma_y^2 = \sigma^2)$$

Now, $r_{uv} = \dfrac{\Sigma uv}{\sqrt{\Sigma u^2\sqrt{\Sigma v^2}}} = \dfrac{0}{n\sigma_u\sigma_v} = 0$

Exercise 8(A)

1. What is meant by correlation ? Does correlation signify the existence of cause and effect relationship between the two variables ?

2. What do you mean by correlation ? What are the different types of correlation ? How are the correlation between two variables measured ?

Website :- https://www.arjun00.com.np

3. Define Karl–Pearson's correlation coefficient. Discuss the relative merits and demerits of Karl Pearson's correlation coefficient.

4. What is the coefficient of correlation between two variables ? Prove that the correlation coefficient between two variables is independent of change of origin and scale.

5. What is scatter diagram ? How does it help to give the idea about the correlation between two variables ?

6. Define correlation coefficient between two variables. Prove that the correlation coefficient between two variables lies between -1 and 1.

7. Draw a scatter diagram from the following data:

Ht.(inch) :	62	72	70	60	67	70	64	65	60	70
Wt.(lbs) :	50	65	63	52	56	60	59	58	64	65

Also indicate whether correlation is positive or negative.

8. If the covariance between the variables X and Y is 18 and the variances of X and Y are 16 and 81 respectively, find the coefficient of correlation between them.

9. Calculate the coefficient of correlation between X and Y series from the following data:

	Series X	SeriesY
No. of observations	15	15
s. d.	3.01	3.03

$$\Sigma(X - \overline{X})(Y - \overline{Y}) = 122$$

10. Calculate the coefficient of correlation using product moment formula from the data of height and weight given below:

Ht. (cms):	160	162	165	161	163	164	165
Wt. (kgs):	63	62	64	63	62	66	68

11. Calculate Karl Pearson's correlation coefficient between the ages of the husbands x and wives y from the following data taking 24 and 22 as the assumed means of x and y respectively.

x :	23	22	24	23	26	27	28	20	30	20
y :	20	18	20	21	21	22	24	23	25	26

Website :- https://www.arjun00.com.np

12. Find Karl Pearson's coefficient of correlation between sales and expenses of the following eight firms.

Firm :	1	2	3	4	5	6	7	8
Sales :	50	55	55	60	65	65	65	60
Expenses :	11	13	14	16	16	15	15	14

13. From the following table calculate the coefficient of correlation by Karl Pearson's method.

X :	6	2	10	4	8
Y :	9	11	?	8	7

Arithmetic means of X and Y series are 6 and 8 respectively.

14. A student calculates the values of r as 0.81 when the value of n is 10 and concludes that r is highly significant. Is his conclusion correct ?

Also determine the limits of population correlation coefficient.

15. Find the coefficient of correlation between the age and playing habit of the following students.

Age(years) :	15	16	17	18	19	20
No. of students :	125	100	75	60	50	40
No. of regular players :	100	75	45	24	15	6

16. In order to find the correlation coefficient between two variables X and Y from 12 pair of observations, the following calculations were made :

$\Sigma X = 30,\ \Sigma Y = 5,\ \Sigma X^2 = 670,\ \ \Sigma Y^2 = 285,\ \Sigma XY = 334.$

On subsequents verification it was found that the pair (X = 11, Y = 4) was copied wrongly, the correct value being (X = 10, Y = 14). Find the correct value of correlation coefficient.

17. Show that if x' , y' are the deviations of the variables x, y from their means

$$r = 1 - \frac{1}{2N} \Sigma f \left(\frac{x'}{\sigma x} - \frac{y'}{\sigma y}\right)^2,$$

$$r = -1 + \frac{1}{2N} \Sigma f \left(\frac{x'}{\sigma x} + \frac{y'}{\sigma y}\right)^2$$

Deduce that $-1 \le r \le 1$.

Website :- https://www.arjun00.com.np

18. If x and y are two uncorrelated variables and if $u = x + y$ and $v = x - y$ find the correlation between u and v in terms of σx and σy, the s. d. 's of x and y respectively.

19. Show that the coefficient of correlation r between two variables x and y is given by

$$r = \frac{(\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2)}{2\sigma_x\sigma_y}$$

where σ_x^2, σ_y^2 and σ_{x-y}^2 are the variances of x, y and x - y respectively.

20. Show that if a and b are constants and r is the correlation between x and y then the correlation between ax and by is equal to r if the signs of a and b are alike and to -r if they are different. Also show that if the constants a, b, c are positive, correlation between ax + by and cy is equal to

$$\frac{(ar\sigma_x + b\sigma_y)}{\sqrt{a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma_y}}$$

21. Family income and its percentage spent on food in the case of hundred families gave the following bivariate frequency distribution. Calculate the coefficient of correlation.

Food Expd (%)	Family income (Rs.)				
	200-300	300-400	400-500	500-600	600-700
10-15	-	-	-	3	7
15-20	-	4	9	4	3
20-25	7	6	12	5	-
25-30	3	10	19	8	-

Answers

8. 0.5 9. 0.89 10. 0.725 11. 0.198 12. 0.855

13. 0.92 14. yes; 0.701, 0.919 15. -0.991 16. 0.775

18. $\frac{\sigma_x^2 - \sigma_y^2}{\sigma_x^2 + \sigma_y^2}$ 21. -0.438

Website :- https://www.arjun00.com.np

8.9 Line Regression

The regression analysis is used to determine the strength (i.e., degree) of the relationship (or correlation) between two variables. The word regression is stepping (or returning) back to the original position. It is a statistical device. It helps to estimate (or predict) the value of one variable if the value of other is known. The unknown variable is called dependent and that of the variable whose value is known is the independent variable. The average relationship between two variables is known as simple linear regression analysis.

8.10 Regression Equations and Regression Coefficients

The regression lines, given by two variables (independent and dependent), which are expressed interms of algebraic relations are known as regression equations. There are two types of regression lines (equations): (a) The regression equation of y on x and (b) The regression equation of x on y.

(a) **Regression equation of y on x:** The regression equation of y on x is

$$y = a + bx \tag{1}$$

where a and b are constants to be determined. The values of a and b can be obtained by solving the equations

$$\Sigma y = na + b\Sigma x \tag{2}$$

$$\text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2 \tag{3}$$

For the values of a and b the equation (1) gives the required line of regression of y on x where b is the regression coefficient of y on x denoted by b_{yx}. It gives the best estimate, of y for the given value of x.

(b) **Regression equation of x on y:** The regression equation of x on y is

$$x = a + by \tag{2}$$

Website :- https://www.arjun00.com.np

where a and b are constants to be determined. The values of a and b can be obtained by solving the equations

$$\Sigma x = na + b\Sigma y \quad (4)$$

$$\text{and} \quad \Sigma xy = a\Sigma y + b\Sigma y^2 \quad (5)$$

For the values of a and b the equations (2) gives the required line of regression of x on y where b is the regression coefficient of x on y. It is denoted by b_{xy}. It gives the best estimate of x for the given value of y.

Notes:

(i) If two lines of regression coincide with each other then there is perfect correlation between them.

(ii) If two lines of regression intersect at right angle then there is no correlation between them.

(iii) Nearer the two lines of regression implies the high degree of correlation and farther (or more wider between) two lines of regression implies the low degree of correlation.

Website :- https://www.arjun00.com.np

EXAMPLE 1:

Obtain the two lines of regression from the following data :

x :	10	12	13	17	18
y :	5	6	7	9	13

Solution :

CALCULATION FOR REGRESSION EQUATIONS

x	y	xy	x^2	y^2
10	5	50	100	25
12	6	72	144	36
13	7	91	169	49
17	9	153	289	81
18	13	234	324	169
$\Sigma x = 70$	$\Sigma y = 40$	$\Sigma xy = 600$	$\Sigma x^2 = 1026$	$\Sigma y^2 = 360$

Let the regression equation of y on x be

$$y = a + bx \quad .....(i)$$

To find the values of a and b we have the following two normal equations.

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Substituting the values of n, Σx, Σy, Σxy, Σx^2 in (ii) & (iii), we have

$$40 = 5a + 70b \quad .....(ii)$$

$$600 = 70a + 1026b \quad .....(iii)$$

Website :- https://www.arjun00.com.np

Solving (ii) and (iii), a = -4.17, b = 0.87

Substituting the values of a and b in (i), the regression equation of y on x is

$$y = -4.17 + 0.87x$$

Again, let the regression equation of x on y be

$$x = a' + b'y \quad ....(iv)$$

Then, we have the following two normal equations.

$$\Sigma x = na' + b'\Sigma y$$

$$\Sigma xy = a'\Sigma y + b'\Sigma y^2$$

Substituting the values

$$70 = 5a' + 40b' \quad .....(v)$$

$$600 = 40a' + 360b' \quad .....(vi)$$

Solving (v) and (vi)

$$a' = 6, \ b' = 1$$

Substituting the values of a' and b' in (iv), the regression equation of x on y is $x = 6 + y$

Alternative Method

In the above method, we find the regression equations directly from actual data. The regression equations can easily be obtained when the deviations of the items are taken from the assumed mean. The regression equations in such cases would be as follows.

The equations of lines of regression of y on x and x on y are

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

and $$x - \overline{x} = b_{xy}(y - \overline{y})$$ respectively.

where $\overline{x}$ and $\overline{y}$ are the arithmetic means of x and y -series and b_{yx}, b_{xy} the regression coefficients of y on x and x on y respectively and are given by

$$b_{yx} = \frac{\Sigma(x - \overline{x})(y - \overline{y})}{\Sigma(x - \overline{x})^2}$$

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

Website :- https://www.arjun00.com.np

and $$b_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

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

σ_x and σ_y represent the s.d of x and y -series respectively and r, the correlation coefficient between x and y.

Also, $$b_{yx} = \frac{n\Sigma uv - \Sigma u . \Sigma v}{n\Sigma u^2 - (\Sigma u)^2}$$

and $$b_{xy} = \frac{n\Sigma uv - \Sigma u . \Sigma v}{n\Sigma v^2 - (\Sigma v)^2}$$

where $u = x - a$ and $v = y - b$.

Website :- https://www.arjun00.com.np

EXAMPLE 2 :

Calculate the regression equations of y on x and x on y from the data given below.

Price(Rs) :	10	12	13	12	16	15
Amount demanded :	40	38	43	45	37	43

Solution :

CALCULATION FOR REGRESSION EQUATIONS

Price(x)	Demand (y)	u = x - 13	v = y - 40	u^2	v^2	uv
10	40	-3	0	9	0	0
12	38	-1	-2	1	4	2
13	43	0	3	0	9	0
12	45	-1	5	1	25	-5
16	37	3	-3	9	9	-9
15	43	2	3	4	9	6
		$\sum u = 0$	$\sum v = 6$	$\sum u^2 = 24$	$\sum v^2 = 56$	$\sum uv = -6$

$$\bar{x} = a + \frac{\sum u}{n} = 13 + \frac{0}{6} = 13$$

$$\bar{y} = b + \frac{\sum v}{n} = 40 + \frac{6}{6} = 41$$

Website :- https://www.arjun00.com.np

$$b_{yx} = \frac{n\Sigma uv - \Sigma u.\Sigma v}{n\Sigma u^2 - (\Sigma u)^2} = \frac{6 \times (-6) - 0 \times 6}{6 \times 24 - (0)2}$$

$$= \frac{-6 \times 6}{6 \times 24} = -0.25$$

and, $$b_{yx} = \frac{n\Sigma uv - \Sigma u.\Sigma v}{n\Sigma v^2 - (\Sigma v)^2} = \frac{6 \times (-6) - 0 \times 6}{6 \times 56 - (6)^2}$$

$$= \frac{-6 \times 6}{6 \times 50} = -0.12$$

Now the regression equation of y on x is

$$y - \overline{y} = b_{yx} \ (x - \overline{x})$$

or, $$\overline{y} - 41 = -0.25\,(x - 15)$$

$$\therefore \quad y = -0.25x + 44.25 \text{ .....(i)}$$

Again the regression equation of x on y is

$$x - \overline{x} = b_{xy} \ (y - \overline{y})$$

or, $$x - 13 = -0.12\,(y - 41)$$

$$\therefore \quad x = -0.12y + 17.92 \text{ .....(ii)}$$

Website :- https://www.arjun00.com.np

8.8.4 Properties of regression coefficients

1. The correlation coefficient is the geometric mean between two regression coefficients.

σ_x = standard deviation of the variate x

σ_y = standard deviation of the variate y

r = correlation coefficient between x and y

b_{yx} = regression coefficient of y on x

b_{xy} = regression coefficient of x on y

Then $b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r \frac{\sigma_x}{\sigma_y}$

Now, $b_{yx} \cdot b_{xy} = r \frac{\sigma_y}{\sigma_x} \cdot r \frac{\sigma_x}{\sigma_y} = r^2$

$\therefore \quad r = \pm \sqrt{b_{yx} \cdot b_{xy}}$

which shows that r is the geometric mean between b_{yx} and b_{xy}.

Website :- https://www.arjun00.com.np

2. Both regression coefficients must have the same sign. The sign of the correlation coefficient is same as that of regression coefficient.

3. The product of two regression coefficient must be less than or equal to 1.

Since, $b_{yx} \cdot b_{xy} = r^2$ and $r \leq 1$

so, $b_{yx} \cdot b_{xy} \leq 1$.

4. Regression coefficients are independent of change of origin but not of scale.

Let the variables x and y be changed to u and v respectively by the following relation :

$$u = \frac{x - a}{h} \quad \text{and} \quad v = \frac{y - b}{k}$$

where a, b, h and k are constants.

we know that, $r_{xy} = r_{uv}$, $\sigma_x = h\sigma_u$ and $\sigma_y = k\sigma_v$

So, $$b_{yx} = r_{xy}\frac{\sigma_y}{\sigma_x} = r_{uv}\frac{k\sigma_v}{h\sigma_u}$$

$$\therefore \quad b_{yx} = \frac{k}{h} r_{uv} \frac{\sigma_v}{\sigma_u}$$

Similarly, $b_{xy} = \frac{h}{k} r_{uv} \frac{\sigma_u}{\sigma_v}$

Thus, the regression coefficient are independent of origin but of scale.

Website :- https://www.arjun00.com.np

EXAMPLE 1 :

From the following data, find the equations of two lines of regression and calculate the expected average height of the son when the height of the father is 167.5 cms.

Ht. of father (cm) :	165	166	167	167	168	169	171	173
Ht. of son (cm) :	167	168	164	168	172	170	169	170

Also , calculate the correlation coefficient between the heights of the father and the son.

Solution :

Let x and y be the heights of father and son respectively.

COMPUTATION OF REGRESSION EQUATIONS

x	y	u = x - 168	v = y - 168	u^2	v^2	uv
165	167	-3	-1	9	1	3
166	168	-2	0	4	0	0
167	164	-1	-4	1	16	4
167	168	–1	0	1	0	0
168	172	0	4	0	16	0
169	170	1	2	1	4	2
171	169	3	1	9	1	3
173	170	5	2	25	4	10
		$\sum u = 2$	$\sum v = 4$	$\sum u^2 = 50$	$\sum v^2 = 42$	$\sum uv = 22$

Website :- https://www.arjun00.com.np

Now, $\overline{x} = a + \frac{\Sigma u}{n} = 168 + \frac{2}{8} = 168.25$

and $\overline{y} = b + \frac{\Sigma v}{n} = 168 + \frac{4}{8} = 168.50$

Again, $b_{yx} = \frac{n\,\Sigma uv - \Sigma u \,.\, \Sigma v}{n\,\Sigma u^2 - (\Sigma u)^2} = \frac{8 \times 22 - 2 \times 4}{8 \times 50 - 4}$

$= \frac{176 - 8}{400 - 4} = \frac{168}{396} = 0.424$

and, $b_{xy} = \frac{n\Sigma uv - \Sigma u \,.\, \Sigma v}{n\,\Sigma v^2 - (\Sigma v)^2} = \frac{8 \times 22 - 2 \times 4}{8 \times 42 - 16}$

$= \frac{176 - 8}{336 - 16} = \frac{168}{320} = 0.525$

The equation of the line of regression of y on x is

$$y - \overline{y} = b_{yx}\,(x - \overline{x})$$

or, $y - 168.5 = 0.424\,(x - 168.25)$

$\therefore$ $y = 0.424x + 97.162$ .....(i)

Again the equation of line of regression of x on y is

$$x - \overline{x} = b_{xy}\,(y - \overline{y})$$

or, $x - 168.25 = 0.525\,(y - 168.5)$

$\therefore$ $x = 0.525y + 79.7875$ .....(ii)

For estimation of height of the son, put $x = 167.5$ in (i)

then, $y = 0.424 \times 167.5 + 97.162 = 168.182$ cms

$\therefore$ The height of the son = 168.182 cms.

Also, $r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{0.424 \times 0.525} = 0.472$

Website :- https://www.arjun00.com.np

EXAMPLE 2

Find the most likely production corresponding to a rainfall 40 mm, from the following data :

	Rainfall (mm.)	Production (m. ton)
Average :	30	500
s.d. :	5	100

Coefficient of correlation = 0.8

Solution :

Let rainfall and production be denoted by x and y respectively.

Then, $\overline{x} = 30$, $\overline{y} = 500$, $\sigma_x = 5$, $\sigma_y = 100$, $r = 0.8$

First we find the regression equations of y on x (as the value of y is required).

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.8 \times \frac{100}{5} = 16$$

The regression equations of y on x is

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 500 = 16(x - 30)$$

$$\therefore \quad y = 16x + 20, \text{.....(i)}$$

For the estimation of production, put x = 40 in (i)

$$y = 16 \times 40 + 20 = 660$$

$\therefore$ The volume of production corresponding to the rainfall 40mm. is 660m. tons.

Website :- https://www.arjun00.com.np

EXERCISE 8 (C)

4. *Find the regression equations of y on x and x on y from the following data and estimate the probable value of x when y is 17.*

x :	5	9	13	17	21
y :	3	8	13	18	23

Solution:

x	y	u= x–13	v= y–13	u^2	v^2	uv
5	3	–8	–10	64	100	80
9	8	–4	–5	16	25	20
13	13	0	0	0	0	0
17	18	4	5	15	25	20
21	23	8	10	65	100	80
		$\Sigma u = 0$	$\Sigma v = 0$	$\Sigma u^2=160$	$\Sigma v^2=250$	$\Sigma uv=200$

$$\overline{x} = a + \frac{\Sigma u}{n} = 13 + \frac{0}{5} = 13$$

$$\overline{y} = a + \frac{\Sigma v}{n} = 13 + \frac{0}{5} = 13$$

$$b_{yx} = \frac{n\Sigma uv - \Sigma u.\Sigma v}{n\Sigma u^2 - (\Sigma u)^2} = \frac{5 \times 200 - 0}{5 \times 160 - 0} = 1.25$$

$$b_{xy} = \frac{n\Sigma uv - \Sigma u.\Sigma v}{n\Sigma v^2 - (\Sigma v)^2} = \frac{5 \times 200 - 0}{5 \times 250 - 0} = 0.8$$

The regression equation of y on x is.

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 13 = 1.25(x - 13)$$

$$\therefore \quad y = 1.25x - 3.25$$

The regression equation of x on y is

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

$$x - 13 = 0.8(y - 13)$$

$$\therefore \quad x = 0.8y + 2.6$$

When $y = 17$, $x = 0.8y + 2.6 = 0.8 \times 17 + 2.6 = 16.2$

5. *From the following table showing age of cars of a certain make and annul maintenance costs, obtain the regression equation for cost related to age :*

Website :- https://www.arjun00.com.np

Age of cars (years) :	2	4	6	8	10	12	14
Maintenance cost ('000' Rs)	2	5	6	7	10	11	15

Also calculate the maintenance cost of the car 20 years old.

Solution:

Let x and y represent age of cars and y, the maintenance cost.

x	y	u=x−8	v=y−7	u^2	v^2	uv
2	2	−6	−5	36	25	30
4	5	−4	−2	16	4	8
6	6	−2	−1	4	1	2
8	7	0	0	0	0	0
10	10	2	3	4	9	6
12	11	4	4	16	16	16
14	15	6	8	36	64	48
		$\Sigma u = 0$	$\Sigma v = 7$	$\Sigma u^2=112$	$\Sigma v^2=119$	$\Sigma uv=110$

$$\overline{x} = a + \frac{\Sigma u}{n} = 8 + 0 = 8$$

$$\overline{y} = a + \frac{\Sigma v}{n} = 7 + \frac{7}{7} = 8$$

$$b_{yx} = \frac{n\Sigma uv - \Sigma u.\Sigma v}{n\Sigma u^2 - (\Sigma u)^2} = \frac{7 \times 110 - 0}{7 \times 112 - 0} = 0.98$$

The regression equation of y on x is

$$y - \overline{y} = b_{yx}(x - \overline{y})$$

or, $y - 8 = 0.98(x - 8)$

$\therefore$ $y = 0.98 + 0.16$

When $x = 20$, $y = 0.98 \times 20 + 0.16$

$= \text{Rs. } 19.76$ thousands

6. *The following table shows the test scores made by salesman on the intelligence test and weekly sales.*

Sales man :	1	2	3	4	5	6	7	8	9	10
Test scores :	40	70	50	60	80	50	90	40	60	60
Sales (units) :	2.5	6.0	4.5	5.0	4.5	2.0	5.5	3.0	4.5	3.0

Calculate the regression line of sales on test scores and estimate the most probable weekly sales if a salesman makes a score of 70.

Website :- https://www.arjun00.com.np

Solution:

Let x and y represent test score and sales.

x	y	u=(x−60)/10	v=(y−3.5)/.5	u^2	v^2	uv
40	2.5	−2	−2	4	4	4
70	6	1	5	1	25	5
50	4.5	−1	2	1	4	−2
60	5	0	3	0	9	0
80	4.5	2	2	4	4	4
50	2	−1	−3	1	9	3
90	5.5	3	4	9	16	12
40	3	−2	−1	4	1	2
60	4.5	0	2	0	4	0
60	3	0	−1	0	1	0
		$\Sigma u = 0$	$\Sigma v = 11$	$\Sigma u^2 = 24$	$\Sigma v^2 = 77$	$\Sigma uv = 28$

$$\overline{x} = a + \frac{\Sigma u}{n} = 60 + 0 = 60$$

$$\overline{y} = a + \frac{\Sigma v}{n} \times h = 3.5 + \frac{11}{10} \times .5 = 4.05$$

$$b_{yx} = \frac{k}{h} \frac{n\Sigma uv - \Sigma u.\Sigma v}{n\Sigma u^2 - (\Sigma u)^2}$$

$$= \frac{0.5}{10} \times \frac{10 \times 28}{10 \times 24} = 0.058$$

The regression equation of y on x is

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 4.05 = 0.058\,(x - 60)$$

$$\therefore \quad y = 0.058x + 0.57$$

When $x = 70$, $y = 0.058 \times 70 + 0.57 = 4.63$

7. *From the data given below find,*

(a) coefficient of correlation between the ages of husbands and wives (b) the regression equations (c) the most likely age of husbands when wife's age is 25.

Age of husb : [x]	22	23	23	24	26	27	27	28	30	30
Age of wife : [y]	18	20	21	20	21	22	23	24	25	26

Website :- https://www.arjun00.com.np

Solution:

x	y	u=x−26	v=y−22	u^2	v^2	uv
22	18	−4	−4	16	16	16
23	20	−3	−2	9	4	6
23	21	−3	−1	9	1	3
24	20	−2	−2	4	4	4
26	21	0	−1	0	1	0
27	22	1	0	1	0	0
27	23	1	1	1	1	1
28	24	2	2	4	4	4
30	25	4	3	16	9	12
30	26	4	4	16	15	16
		$\Sigma u = 0$	$\Sigma v = 0$	$\Sigma u^2=76$	$\Sigma v^2=56$	$\Sigma uv=62$

(a) $$r = \frac{n\Sigma uv - \Sigma u.\Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{10 \times 62 - 0}{\sqrt{10 \times 76 - 0}\sqrt{10 \times 56 - 0}}$$

$$= \frac{10 \times 62}{\sqrt{760}\sqrt{560}} = \frac{620}{652.38} = 0.95$$

(b) $$\overline{x} = a + \frac{\Sigma u}{n} = 26 + 0 = 26$$

$$\overline{y} = a + \frac{\Sigma v}{n} = 22 + 0 = 22$$

$$b_{yx} = \frac{n\Sigma uv - \Sigma u.\Sigma v}{n\Sigma u^2 - (\Sigma u)^2} = \frac{10 \times 62 - 0}{10 \times 76 - 0} = 0.816$$

$$b_{xy} = \frac{n\Sigma uv - \Sigma u.\Sigma v}{n\Sigma v^2 - (\Sigma v)^2} = \frac{10 \times 62 - 0}{10 \times 56.0} = 1.107$$

The regression equation of y on x is

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 22 = 0.816(x - 26)$$

$$\therefore \quad y = 0.816x + 0.784$$

The regression equation of x on y is

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

Website :- https://www.arjun00.com.np

$$x - 26 = 1.107\,(y - 22)$$

$$\therefore \quad x = 1.107y + 1.646$$

(c) When $y = 25$, $x = 1.107 \times 25 + 1.646 = 29.321$

8. *You are given below the following information about advertising expenditure and sales.*

	Advertising (X) Expenditure (Rs. Lakhs)	Sales (Y) (Rs. Lakhs)
Mean :	10	90
s . d. :	3	12

coefficient of correlation $= 0.8$

(a) *Calculate the two regression lines*

(b) *Find the likely sales when advertising expenditure is Rs. 15 lakhs*

(c) *What should be the advertising expenditure if the company wants to attain sales target of Rs. 120 lakhs ?*

Solution:

$\overline{x} = 10$, $\overline{y} = 90$, $\sigma_x = 3$, $\sigma_y = 12$, $\sigma = 0.8$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.8 \times \frac{12}{3} = 3.2$$

$$b_{xy} = r\frac{\sigma_y}{\sigma_y} = 0.8 \times \frac{3}{12} = 0.2$$

(a) The regression equation of y on x is

$$y - \overline{y} = b_{yx}\,(x - \overline{x})$$

$$y - 90 = 3.2\,(x - 10)$$

$$\therefore \quad y = 3.2x + 58$$

The regression equation of x on y is

$$x - \overline{x} = b_{xy}\,(y - \overline{y})$$

$$x - 10 = 0.2\,(y - 90)$$

$$\therefore \quad x = 0.2y - 8$$

(b) When $x = 15$, $y = 3.2 \times 15 + 58 =$ Rs. 106 lakhs

(c) When $y = 120$, $x = 0.2 \times 120 - 8 =$ Rs. 16 lakhs

9. *The following table gives the age of the persons and their blood pressures :*

Age (yrs) x :	66	38	56	42	72	36	63	47	55	45
Blood pressure y :	145	124	147	125	160	118	149	128	150	124

Website :- https://www.arjun00.com.np

Estimate the blood pressure of a man whose age is 50 years. Also calculate the value of r.

Solution:

x	y	u=x–47	u=y–128	u^2	v^2	uv
66	145	19	17	361	289	323
38	124	–9	–4	81	16	36
56	147	9	19	81	361	171
42	125	–5	–3	25	9	15
72	160	25	32	625	1024	800
36	118	–11	–10	121	100	110
63	149	16	21	256	441	336
47	128	0	0	0	0	0
55	150	8	22	64	484	176
45	124	–2	–4	4	16	8
		$\Sigma u=50$	$\Sigma v=90$	$\Sigma u^2=1618$	$\Sigma v^2=2740$	$\Sigma uv=1975$

$$\overline{x} = a + \frac{\Sigma u}{n} = 47 + \frac{50}{10} = 52$$

$$\overline{y} = a + \frac{\Sigma v}{n} = 128 + \frac{90}{10} = 137$$

$$b_{yx} = \frac{n\Sigma uv - \Sigma u.\Sigma v}{n\Sigma u^2 - (\Sigma u)^2} = \frac{10 \times 1975 - 50 \times 90}{10 \times 1618 - (50)^2}$$

$$= \frac{15250}{13680} = 1.115$$

The regression equation of y on x

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 137 = 1.115(x - 52) \quad \therefore y = 1.115x + 79.02$$

when $x = 50$, $y = 1.115 \times 50 + 79.02 = 134.77$

Again, $$b_{xy} = \frac{n\Sigma uv - \Sigma u.\Sigma v}{n\Sigma v^2 - (\Sigma v)^2} = \frac{10 \times 1975 - 50 \times 90}{10 \times 2740 - (90)^2}$$

$$= \frac{15250}{19300} = 0.79$$

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{1.115 \times 0.79} = 0.94$$

Website :- https://www.arjun00.com.np

10. *Given the following data*

Regression equation : $4x - 5y + 33 = 0$,

$20x - 9y - 107 = 0$, variance of $x = 9$

Find (i) The mean values of x and y

(ii) The coefficient of correlation between x and y

(iii) The standard deviation of y

Solution:

(i) Let $\overline{x}$ and $\overline{y}$ be the mean values.

$$4\overline{x} - 5\overline{y} + 33 = 0$$

$$20\overline{x} - 9\overline{y} - 107 = 0$$

Solving these two equations. $\overline{x} = 13$, $\overline{y} = 17$

(ii) $4x - 5y + 33 = 0 \quad \therefore y = \frac{4}{5}x + \frac{33}{5}$

$$b_{yx} = \frac{4}{5}$$

$$20x - 9y - 107 = 0 \quad \therefore x = \frac{9}{20}y + \frac{107}{20}$$

$$b_{xy} = \frac{9}{20}$$

$$r = \sqrt{b_{yx}.b_{xy}} = \sqrt{\frac{4}{5} \times \frac{9}{20}} = 0.6$$

(iii) $b_{yx} = \frac{4}{5}$ or, $r\frac{\sigma y}{\sigma x} = \frac{4}{5}$ or, $0.6 \times \frac{\sigma y}{3} = \frac{4}{5}$

$\therefore \sigma y = 4$

11. *Given the following data calculate the mark in Mathematics obtained by a students who has scored 60 marks in English :*

Mean mark in Math = 80, Mean mark in English = 50

s.d. ,, ,, = 15, s.d. ,, ,, = 10

r between Math and English = -0.4

Solution:

Suppose $\overline{x}$ and $\overline{y}$ are mean marks in Math and English.

$\overline{x} = 80$, $\overline{y} = 50$

Website :- https://www.arjun00.com.np

$\sigma x = 15, \sigma y = 10, r = -0.4$

$$b_{xy} = r\frac{\sigma x}{\sigma y} = -0.4 \times \frac{15}{10} = -0.6$$

The regression equation of x on y is

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

$$x - 80 = -0.6(y - 50)$$

$$\therefore \quad x = -0.6y + 110$$

when $y = 60$, $x = -0.6 \times 60 + 110 = 74$

12. *If $a_1x + b_1y + c_1 = 0$ and $a_2x + b_2y + c_2 = 0$ are the equations of lines of regression of y on x and x on y respectively. Find the correlation coefficient between the two variables x and y. Also prove that $a_1b_2 \le a_2b_1$.*

Solution:

$$a_1x + b_1y + c_1 = 0$$

$$b_1y = -a_1x - c_1$$

$$\therefore \quad y = -\frac{a_1}{b_1}x - \frac{c_1}{b_1}$$

$$\therefore \quad b_{yx} = -\frac{a_1}{b_1}$$

Again, $a_2x + b_2y + c_2 = 0$

or, $a_2x = -b_2y - c_2$

$$\therefore \quad x = -\frac{b_2}{a_2}y - \frac{c_2}{a_2}$$

$$\therefore \quad b_{xy} = -\frac{b_2}{a_2}$$

$$r^2 = b_{yx}.b_{xy} = -\frac{a_1}{b_1}\left(-\frac{b_2}{a_2}\right) = \frac{a_1b_2}{a_2b_1}$$

$$\therefore \quad r = \pm\sqrt{\frac{a_1b_2}{a_2b_1}}$$

But $r^2 \le 1$

so, $\frac{a_1b_2}{a_2b_1} \le 1$

$\therefore \quad a_1b_2 \le a_2b_1$

Website :- https://www.arjun00.com.np

Chapter 9

PROBABILITY

9.1 Introduction

The world "Probability" is the synonym of the words "most likely", "probably", "almost certain", "chance" which we may consciously or unconsciously use in our daily life by saying that "probably you may get promotion this time" "it is almost certain that he may pass the examination" etc. All the terms mentioned above related to the uncertainity. A numerical measure of uncertainity is provided by a very important branch of statistics known as the "theory of probability." This theory originated by a gambler Chevalier de Mere in 17th century, had its beginning with games of chance such as throwing a dice, tossing of a coin, drawing a card and so on. But now a days, it has wide application in economics, business, social sciences and even in our daily life too.

9.2 Basic definitions

Before giving the definition of probability, we explain some terms which are used in the definition of probability.

Experiment

The processes which when performed, result different possible outcomes or cases are known as the experiment. While performing an experiment repeatedly under the same condition if the result obtained is not unique, but may be any one of the possible outcomes, then the experiment is known as a random experiment

Trial and Event

Performing of a random experiment is called a trial and outcome or a combination of outcomes of an experiment is termed as an event.

Drawing card from a deck of 52 cards, is a trial or an experiment and getting any one of the cards is an event. An event is said to be a sure event if its occurrence is certain and an event which can never occur is said to be an impossible event. We use the term "success" whenever an event of an experiment under consideration, takes place and fallure whenever it does not.

Website :- https://www.arjun00.com.np

There are two types of events (i) simple (ii) compound. An event is said to be simple if it relates to the occurrence or non-occurrence of a single event. On the other hand, when two or more events occur in connection with each other, their simultaneous occurrence is called a compound event.

Getting a head or a tail when a coin is tossed is simple event. The event that "the sum of two numbers shown on the upper faces of two unbaised dice is seven when the two dice are thrown simultaneously is a compound event.

Exhaustive cases

The number of cases which include all possible outcomes of a random experiment is said to be the exhaustive cases for the experiment. In throwing a die, the turning up of 1, 2, 3, 4, 5 and 6 marked in a die made exhaustive cases. Thus the total number of exhaustive cases in throwing a die is 6.

Equally likely cases

While performing an experiment if any one of the possible outcomes may occur but no one case can be expected to occur more than the other then the cases are said to be equally likely.

If a die is rolled, any one of the six numbers marked in the faces of a die, may turn up. So there are 6 equally likely cases in throwing a die.

Mutually exclusive cases (events)

Two or more events are said to be mutually exclusive if their simultaneous occurence is not possible. If a coin is tossed either head or tail will occur, so head and tail are two mutually exclusive events.

Favourable cases

The cases or the outcomes of a random experiment which entail the happening of an event are known as the cases favourable to that event.

In throwing a die, the cases favourable to "getting an odd numbers" are 3.

Independent cases (Events)

Two events are said to be independent if the occurrence of one event does not effect the occurrence of the other. If a coin and a die are thrown, the turning head up in a coin will not effect the getting 1 on the die.

Website :- https://www.arjun00.com.np

9.3 Permutation and Combination

Permutation

Permutation of a set of objects means the arrangement of objects in some order. Below, we list some of the important formula without their proofs.

(a) The total number of permutation of a set of n different objects taken r at a time denoted by P (n, r) or ${}^nP_r$ is given by

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

where n ! = factorial n = 1, 2, 3..... n.

In particular, 5! = the continued product from 1 to 5

$= 1.2.3.4.5$

$= 5 \times 4! = 5 \times 4 \times 3!$

Also ${}^np_n = n!$ $(0! = 1)$

(b) The number of permutations of a set of n objects taken all of them at a time where p of them are of one kind, q of them the second kind, r of them of the third kind is

$$ \frac{n!}{p!\ q!\ r!} $$

Combination

Combination of a set of objects meant the selection of objects without regard to any order of arrangement.

(a) The total number of selections of a set of n different objects taken r at a time denoted by C (n, r) or ${}^nC_r$ is given by

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

(b) $ {}^nC_r = {}^nC_{n-r} $

9.4 Classical or Priori Definition of Probability

If there are n exhaustive, mutually exclusive and equally likely cases and m of them are favourable to an event E, then the probability of the happening of an event E denoted by P (E) is defined by

$$ P(E) = \frac{m}{n} $$

Website :- https://www.arjun00.com.np

The probability P (E) of happening of an event E satisfies the following property

$$0 \leq P(E) \leq 1.$$

This definition is also known as 'Mathematical definition of probability.

Cor. 1 (i) If E is an impossible event then P (E) = 0

(ii) If E is a sure event then P (E) = 1.

Cor. 2 The sum of the probabilities of the occurrence and non-occurrence of an event is unity.

If the probability of occurrence of an event is denoted by P (E), the probability of its non-occurrence is denoted by $P(\overline{E})$

$$\therefore P(E) + P(\overline{E}) = 1$$

Here E and $\overline{E}$, the two mutually exclusive events, the sum of whose probabilities is unity are known as complementary events.

EXAMPLE 1:

250 bulbs were found to be defective in a lot of 1000 bulbs. Find the probability that a bulb selected at random from the lot is a defective one.

Solution :

n = total number of cases = 1000

m = no. of favourable cases = 250

P(E) = Prob of getting a defective bulb = ?

Now, $$P(E) = \frac{m}{n}$$

$$= \frac{250}{1000} = 0.25$$

9.4.1 Limitation of Classical definition of probability

The classical definition of probability suffers from the following limitations :

(i) This definition does not help if the cases are not equally likely.

(ii) Enumeration of all equally likely cases are not practicable

(iii) This definition fails if the total number of possible cases is infinite

Website :- https://www.arjun00.com.np

9.5 Empirical definition of Probability

If an event E occurs m times in n repetitions of a random experiment then the value approached by the relative frequency $\frac{m}{n}$ when n becomes infinitely large is said to be the probability of an event E. Symbolically,

$$P(E) = \lim_{n \to \infty} \frac{m}{n}$$

provided that the limit exists.

Note: If an unbiased coin is tossed, the classical definition of probability states that the probability of turning up of a head is 0.5. So, if we toss the coin 100 times, classical definition suggests that we get heads 50 times. In practice, this may not be true. In practice, we may get 10 heads or 20 heads or 80 heads in 100 throws. So, empirical definition of probability suggests that if a coin is tossed a large number of times say 1000 times we may on the average expect 50% heads and 50% tails. This conclude that the empirical definition of probability approaches the classical definition of probability if the number of experiments be made infinitely large.

9.5.1 Limitation of Empirical definition of probability

The following are the limitations of empirical definition of probability.

(i) In practice, the probability defined as the limit of a relative frequency can never be obtained however large the number of experiments be made.

(ii) Homogeneous and identical conditions can not be maintained while the experiment is performed repeatedly.

9.6 Two basic laws of probability

The following two theorems are the basic laws of probability

(i) Addition theorem

(ii) Multiplication theorem

(i) Addition theorem (Theorem of total probability)

If A and B are two events with their respective probabilities P(A) and P(B) then the probability of occurrence of at least one of these two events denoted by $P(A \cup B)$ is given by

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

Website :- https://www.arjun00.com.np

where $P(A \cap B)$ is the probability of the simultaneous occurrence of the events A and B.

Let n be the total number of equally likely and exhaustive outcomes (or cases) of a random experiment. If u, v and w be the number of cases favourable to the events A, B and $A \cap B$ (common to A and B) respectively, then

$$P(A) = \frac{u}{n}, \quad P(B) = \frac{v}{n} \quad \text{and} \quad P(A \cap B) = \frac{w}{n}$$

Since the outcomes w are common to both events A and B, so the cases favourable to the event $A \cup B$ are $(u + v - w)$

$$\text{Now, } P(A \cup B) = \frac{u + v - w}{n}$$

$$= \frac{u}{n} + \frac{v}{n} - \frac{w}{n}$$

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

Cor 1. If A and B are two mutually exclusive events then

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

Cor 2. If A, B and C are three events, then the probability of occurrence of at least one of these events is

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

If the events are mutually exclusive, then

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

Note 1. The probability of occurrence of at least one of the events A and B can also be written as $P(A+B)$ or $P(A \text{ or } B)$.

EXAMPLE 2.

What is the probability of drawing a heart or an ace from a deck of 52 cards.

Solution:

There are 13 hearts and 4 aces ina deck of 52 cards; One card i.e. an ace of heart being common.

$$P(A) = \text{Prob. of getting a heart} = \frac{m}{n} = \frac{13}{52}$$

$$P(B) = \text{Prob. of getting an ace} = \frac{m}{n} = \frac{4}{52}$$

Website :- https://www.arjun00.com.np

$P(A \cap B)$ = Prob of getting an ace of heart = $\frac{1}{52}$

$P(A \cup B)$ = " " a heart or an ace = ?

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

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{4}{13}$$

(ii) Multiplication theorem (Theorem of compound probability)

If two events A and B are independent, then the probability of their simultaneous occurrence is equal to product of their individual probabilities.

Let n_1 and n_2 the total number of possible cases for events A and B respectively, have m_1 and m_2 as their respective favourable cases, Then,

$$P(A) = \text{Prob of occurrence of an event A} = \frac{m_1}{n_1}$$

$$P(B) = \text{Prob. of occurrence of an event B} = \frac{m_2}{n_2}$$

Since the total number of possible cases for the events A and B are n_1 and n_2 respectively and since we can combine each of n_1 possible cases of the event A with each of the n_2 possible cases of the event B, so the total number of possible cases for their simultaneous occurrence is $n_1.n_2$. In the same way, the total number of favourable cases for the simultaneous occurrence of the events A and B is $m_1.m_2$.

Now, $P(A \cap B)$ = Prob. of simultaneous occurrence of the events A and B

$$= \frac{m_1.m_2}{n_1.n_2} = \frac{m_1}{n_1} \cdot \frac{m_2}{n_2} = P(A).P(B)$$

Cor. 1 If A, B and C are three independent events, then the probability of their simultaneous occurrence $P(A \cap B \cap C)$ is given by

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

Cor. 2 The probability that the event A will not happen is $1 - P(A)$. Therefore the probability that none of the two independent events A and B will happen, is $[1 - P(A)]\,[1 - P(B)]$.

$\therefore$ The probability that least one event will happen

$$= 1 - [1 - P(A)]\,[1 - P(B)].$$

Note 1: The probability of the simultaneous occurrence of two independent events A and B can also be written as P(AB) or P(A and B).

Website :- https://www.arjun00.com.np

EXAMPLE 3:

Ram and Laxman appear in an interview for the same post The probability of Ram's selection is 1/7 and that of Laxman's selection is 14/15. What is the probability that both will be selected ?

Solution :

P(Ram) = 1/7, P(Laxman) = 14/15

Since the events are independent, so

P(Ram and Laxman) = P(Ram). P(Laxman)

$$= \frac{1}{7} \cdot \frac{14}{15} = \frac{2}{15}$$

ILLUSTRATIVE EXAMPLES

EXAMPLE 1:

A die is thrown once. Determine the probability of getting a number less than or equal to 4.

Solution :

There are six faces in a die sand 1, 2, 3, 4, 5 and 6 are marked in the faces. The number of cases favourable to the event of getting a number less than or equal to 4 is 4.

n = total number possible cases = 6

m = no of favourable cases = 4

P(a number ≤ 4) = ?

$$P(\text{a number} \leq 4) = \frac{m}{n} = \frac{4}{6} = \frac{2}{3}$$

EXAMPLE 2:

A bag contains 4 white marbles and red 8 red marbles. If a marble is drawn, what is the probability that is a white marble ?

Solution:

Since there are 4 white and 8 red marbles in the bag, so the total number of marbles in the bag is 4 +8 which is 12.

n = total number of possible cases = 12

m = no. of favourable cases = 4

P(white marble) = ?

$$P(\text{white marble}) = \frac{m}{n} = \frac{4}{12} = \frac{1}{3}$$

Website :- https://www.arjun00.com.np

EXAMPLE 3:

The probability that a boy will get a scholarship is 0.75 and that a girl will get is 0.72. What is the probability that at least one of them will get the scholarship ?

Solution :

P (A) = Prob that a boy gets a scholarship = 0.75

P (B) = ,, ,, a girl ,, ,, = 0.72

P (AB) = ,, ,, both get scholarship = 0.75×0.72

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

$$= 0.75 + 0.72 - 0.75 \times 0.72 = 0.93$$

EXAMPLE 4:

In a single throw of two dice, find the probability of the following events:

(i) odd number on the first die and 6 on the second

(ii) a total of 11.

Solution :

(i) n = Total number of possible cases = $6 \times 6 = 36$.

The favourable cases for the events of getting odd numbers in first die and 6 on the second are shown below.

First die :	1	3	5
Second die:	6	6	6

m = no. of favourable cases = 3

$$P(\text{odd in 1st. die and 6 on second}) = \frac{m}{n} = \frac{3}{36} = \frac{1}{12}$$

(ii) n = Total number of possible cases = $6 \times 6 = 36$

The favourable cases of getting a total of 11 in single throw of two dice are the events (5, 6) and (6, 5).

So, m = no. of favourable cases = 2

P(total of 11) = ?

$$P(\text{total of } 11) = \frac{m}{n} = \frac{2}{36} = \frac{1}{18}$$

Website :- https://www.arjun00.com.np

EXAMPLE 5:

One urn contains 6 white and 10 black balls. A second urn contains 8 white and 12 black balls. One ball is drawn from each urn. What is the probability that the balls drawn are both white.

Solution:

From first urn:

Total number of balls in first urn = 6 + 10 = 16

n = total no. of possible cases = 16

m = no. of favourable cases = 6

P(white) = ?

$$P(\text{white}) = \frac{m}{n} = \frac{6}{16}$$

From second urn:

Total number of balls in second urn = 8 +12 = 20

n =total no. of possible cases = 20

m = no. of favourable cases = 8

P (white) = ?

$$P\ (\text{white}) = \frac{m}{n} = \frac{8}{20}$$

Since the events are independent, so probability of getting both balls white

$$= P(\text{white}) \times P(\text{white}) = \frac{6}{16} \times \frac{8}{20} = \frac{3}{20}$$

EXAMPLE 6:

The personal department of a company has records which show the following analysis of its 200 engineers.

Age	Bachelor degree only	Master degree	Total
under 30	90	10	100
30-40	20	30	50
over 40	40	10	50
Total	150	50	200

Website :- https://www.arjun00.com.np

If one engineer is selected at random from the company, find

(a) the probability that he has only a bachelor degree.

(b) the probability that he has a master degree given that he is over 40.

(c) the probability that he is under 30, given that he has only a bachelor degree.

Solution:

(a) There are 200 engineers, 150 of them having bachelor degree only.

n = total no. of possible cases = 200

m = no of favourable cases = 150

P(bachelor degree only) = ?

$$P(\text{bachelor degree only} = \frac{m}{n} = \frac{150}{200} = 3/4.$$

(b) There are 40 master degree holder engineers.

10 of them having ages above 40

n = total no. of possible cases = 50

m = no. of favourable cases = 10

P(master degree holder with age above 40) = ?

$$P(\text{master degree holder with age above 40}) = \frac{m}{n} = \frac{10}{50} = 1/5$$

(c) There are 150 bachelor degree holder engineers, 90 of them having ages below 30.

n = total no. of possible cases = 150

m = no. of favourable cases = 90

P (bachelor degree holder with age below 30) = ?

$$P(\text{bachelor degree holder with age below 30}) = \frac{m}{n} = \frac{90}{150} = \frac{3}{5}$$

EXAMPLE 7:

A coin is tossed successively three times. Determine the probability of getting

(i) exactly 2 heads (ii) at least 2 heads

(iii) at most two heads.

Website :- https://www.arjun00.com.np

Solution:

A coin has two faces head and tail. So, when a coin is tossed three times, the outcomes obtained in three trials may be as follows:

First toss :	H	H	H	H	T	T	T	T
Second toss:	H	H	T	T	H	H	T	T
Third toss :	H	T	H	T	H	T	H	T

Total numbers of possible cases is 8. Number of cases with exactly two heads, at least two heads and at most two heads are 3, 4 and 7 respectively. So,

[i] P (exactly two heads) = m/n = 3/8

[ii] P (at last two heads) = m/n = 4/8 = 1/2

[iii] P (at most two heads) = m/n = 7/8

EXAMPLE 8:

Five men in a group of 20 are graduates. If 3 men are chosen out of 20 at random, what is the probability that (i) all are graduates (ii) none of them is graduate (iii) at least one being graduates (iv) 1 is graduate and rest two others ?

Solution :

Out of 20, 5 are graduates and rest 15 others. 3 men are to be chosen out of 20, so total number of possible cases is equal to the combination of 3 out of 20 which is $^{20}C_3$.

[i] n = Total no. of possible cases = $^{20}C_3$

m = no. of favourable cases

= the no. of selection of 3 graduates out of 5 = $^{5}C_3$

$$P \text{ (all graduates)} = \frac{m}{n} = \frac{^{5}C_3}{^{20}C_3} = \frac{5 \times 4 \times 3}{20 \times 19 \times 18} = \frac{1}{114}$$

[ii] There are 15 non-graduate from which 3 are to be selected, they can be selected in $^{15}C_3$

$$P \text{ (none is graduate)} = \frac{m}{n} = \frac{^{15}C_3}{^{20}C_3}$$

$$= \frac{15 \times 14 \times 13}{20 \times 19 \times 18} = \frac{91}{228}$$

Website :- https://www.arjun00.com.np

[iii] Since the events at least one being graduate and none of them is graduate, are complementary, so,

$$P\text{ (atleast one graduate)} = 1 - P\text{ (none is graduate)}$$

$$= 1 - \frac{91}{228} = \frac{137}{228}$$

[iv] Since we need 1 graduate and rest two others, so 1 is to be selected from 5 graduates and rest 2 from 15, so

no. of favourable cases $= {}^5C_1 \times {}^{15}C_2$

$$P\text{ (1 graduate and rest two)} = \frac{m}{n} = \frac{{}^5C_1 \times {}^{15}C_2}{{}^{20}C_3}$$

$$= \frac{15 \times 14 \times 13}{20 \times 19 \times 18} = \frac{35}{76}$$

EXAMPLE 9:

The chance that A can solve a certain problem is $\frac{1}{4}$, the chance that B can solve it is $\frac{2}{3}$, find the chance that the problem will be solved if they both try.

Solution :

The problem will be solved means it is solved by at least one of them.

$$P(A) = \frac{1}{4},\ P(B) = \frac{2}{3};\ P(A \text{ or } B) = ?$$

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

$$= \frac{1}{4} + \frac{2}{3} - \frac{1}{4} \times \frac{2}{3} = \frac{9}{12} = \frac{3}{4}$$

Alternatively:

The chance that A will not solve $= 1 - \frac{1}{4} = \frac{3}{4}$

" " " B " " $= 1 - \frac{2}{3} = \frac{1}{3}$

" " " both will not solve $= \frac{3}{4} \times \frac{1}{3} = \frac{1}{4}$

" " that it will be solved $= 1 - \frac{1}{4} = \frac{3}{4}$

Website :- https://www.arjun00.com.np

EXAMPLE 10:

The probability that Nepal wins a football match against Bangladesh is given to be 2/3. If Nepal and Bangladesh play three matches, what is the probability that

(i) Nepal will lose all the three matches (ii) Nepal will win in all three matches (iii) Nepal will win at least one match (iv) One country will win first two matches and rest one by the other.

Solution:

$$P\,(\text{Nepal}) = \frac{2}{3}, \qquad P\,(\text{Bangladesh}) = \frac{1}{3}$$

where P (Nepal) means the probability that Nepal will win the match etc.

(i) The probability that Nepal will lose all the three matches = The probability that Bangladesh will win all three matches

$$= \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{27}$$

(ii) The probability that Nepal will win in all the three matches = P (Nepal). P (Nepal). P (Nepal)

$$= \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} = \frac{8}{27}$$

(iii) The probability that Nepal will win at least one match

$$= 1 - \frac{1}{27} = \frac{26}{27}$$

(iv) The probability that one country will win 2 matches and 1 by the other = P (Nepal). P (Nepal). P (Bangladesh) + P(Bangladesh). P (Bangladesh). P (Nepal)

$$= \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{6}{27} = \frac{2}{9}.$$

EXAMPLE 11:

A candidate is selected for interview for three posts. for the first post there are 3 candidates, for the second there are four and for the third there are 2. What are his chances of getting at least one post ?

Solution:

P (A) = Prob. of being selected in 1st. post $= \frac{1}{3}$

P (B) = ,, ,, 2nd. ,, $= \frac{1}{4}$

Website :- https://www.arjun00.com.np

P (C) = ,, ,, 3rd. ,, $= \frac{1}{2}$

P (at least one post) = 1 – [1 – P(A)] [1 – P(B)] [1 – P(C)]

$$= 1 - \left(1 - \frac{1}{3}\right)\left(1 - \frac{1}{4}\right)\left(1 - \frac{1}{2}\right)$$

$$= 1 - \frac{2}{3} \cdot \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{4}$$

EXAMPLE 12:

Three groups of children contain respectively 3 girls and 1 boy, 2 girls and 2 boys, 1 girl and 3 boys. One child is selected at random from each group. Show that the probability that three selected consist of 1 girl and 2 boys is 13/32.

Solution:

The selection of 1 girl and 2 boys can be made in the following way.

Group I	Group II	Group III
(3 girls & 1 boy)	(2 girls & 2 boys)	(1 girl & 3 boys)
1 Girl	1 Boy	1 Boy
1 Boy	1Girl	1 Boy
1 Boy	1 Boy	1 Girl

Pro. of selecting 1 girl from group I $= \frac{3}{4}$

,, ,, 1 boy ,, ,, II $= \frac{2}{4}$

,, ,, 1 boy ,, ,, III $= \frac{3}{4}$

P (GBB)= Prob. of selecting 1 girl from group I, 1 boy from group II and 1 boy from group III

$$= \frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{9}{32}$$

Similarly, $P(BGB) = \frac{1}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{3}{32}$

and, $P(BBG) = \frac{1}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{1}{32}$

Website :- https://www.arjun00.com.np

Since the events are mutually exclusive, so the probability of selecting 1 girl and 2 boys

$$= \frac{9}{32} + \frac{3}{32} + \frac{1}{32} = \frac{13}{32}$$

EXAMPLE 13:

The letters of the word "concept" are arranged at random. Find the chance that (i) the two c's come together (ii) the two c's do not come together.

Solution :

(i) There are 7 letters in the word "concept". These letters can be arranged in 7!/2! ways Now, we consider two c's as one; so the remaining 6 letters can be arranged in 6! ways,

P (two c's come together) $= \frac{2 \times 6!}{7!} = \frac{2}{7}$

(ii) P (two c's do not come together) $= 1 - \frac{2}{7} = \frac{5}{7}$

Exercise 9(A)

1. Define the following terms:

   [i] Exclusive cases [ii] Equally likely cases [iii] Mutually exclusive events [iv] Favourable cases

2. Explain independent and mutually exclusive events with examples.

3. Define priori and empirical definitions of probability. What are the limitations of priori definition of probability. In what respect does the mathematical definition of probability differ from the empirical definition of probability.

4. State and prove the addition and multiplication theorems of probability for two events.

5. Define priori definition of probability. Prove the addition theorem of probability when the events are mutually exclusive.

6. A die is thrown once. Determine the probability of getting.

   [i] an even number [ii] a number ≥ 3 [iii] a number ≤ 4 [iv] a number <7 [v] a number > 6.

Website :- https://www.arjun00.com.np

7. A bag contains 9 red, 7 white and 4 black balls. A ball is drawn at random. What is the probability of drawing

[i] a red ball [ii] a white ball [iii] not a black ball.

8. A card is drawn at random from a well-shuffled deck of 52 cards. What is the probability that it is

[i] a red card [ii] a spade [iii] a jack

[iv] an ace, a king or a queen [v] a red 8, a red 9 or a red 10. ?

9. From 30 tickets marked with the first 30 numerals, one is drawn at random. Find the chance that [i] it is a multiple of 5 or of 7 [ii] it is a multiple of 3 or of 7 [iii] it is an odd number.

10. In 120 families each of five children, number of males were distributed as follows:

No of male children per family:	0	1	2	3	4	5
No. of families :	4	18	40	36	20	2

Calculate the probability of male birth.

11. If two dice are rolled once, what is the probability of getting [i] a total score of 7 [ii] a total of 5 or 7 [iii] a total of 5 or 3 in the first dice ?

12. A coin is tossed successively three times. Determine the probability of getting

[i] all three heads [ii] head and tail alternately

[iii] at least one head [iv] 2 heads and one tail

13. A class consists of 60 boys and 40 girls. If two students are chosen at random, what will be the probability that [a] both are boys [b] both are girls [c] one boy and one girl ?

14. A lot contains 10 items of which 3 are defective. Three items are chosen from the lot at random one after another without replacement. Find the probability that all the three are defective.

15. There are 3 physicists, 4 mathematicians, 2 statisticians and 1 chemist. A committee of 4 from among them is to be formed. Find the probability that the committee (i) consists of one of each kind (ii) has at least one mathematician (iii) has the chemist as a member and three others.

16. In a group of 100 businessmen, 10 are bachelors and others are married. What is the probability of selecting a sample of 20 business men out of which 5 are bachelors, if the sample is selected at random.

Website :- https://www.arjun00.com.np

17 A problem in statistics is given to three students A, B, C, whose chances of solving it are $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{5}$ respectively. Find the probability that the problem will be solved.

18. The letters of the word "Triangle" are arranged at random. Find the probability that the word so formed (i) starts with T (ii) ends with e (iii) starts with T and ends with e .

19. A husband and wife appear in an interview for the same post. The probability of husband's selection is $\frac{1}{7}$ and that of wife's selection is $\frac{1}{5}$. What is the probability that (i) both of them will be selected (ii) none of them will be selected (iii) at least one of them will be selected.

20. A blind and deaf school has 100 students out of which 60 are having defective eyes and 50 having defective ears. A student is chosen at random What is the probability that the selected students (a) has defective eye and defective ear (b) has defective eye but not defective ear.

Website:- www.arjun00.com.np

Answers

6. (i) $\frac{1}{2}$ (ii) $\frac{2}{3}$ (iii) $\frac{2}{3}$ (iv) 1 (v) 0 7.

7. (i) $\frac{9}{20}$ (ii) $\frac{7}{20}$ (iii) $\frac{4}{5}$ 8. (i) $\frac{1}{2}$ (ii) $\frac{1}{4}$

(iii) $\frac{1}{13}$ (iv) $\frac{3}{13}$ (v) $\frac{3}{26}$ 9. (i) $\frac{1}{3}$ (ii) $\frac{13}{30}$

(iii) $\frac{1}{2}$ 10. $\frac{37}{75}$ 11. (i) $\frac{1}{6}$ (ii) $\frac{5}{18}$ (iii) $\frac{1}{4}$

12. (i) $\frac{1}{8}$ (ii) $\frac{1}{4}$ (iii) $\frac{7}{8}$ (iv) $\frac{3}{8}$ 13. (a) $\frac{{}^{60}C_2}{{}^{100}C_2}$

(b) $\frac{{}^{40}C_2}{{}^{100}C_2}$ (c) $\frac{{}^{60}C_1 \times {}^{40}C_1}{{}^{100}C_2}$ 14. $\frac{1}{120}$ 15. (i) $\frac{4}{35}$

(ii) $\frac{13}{14}$ (iii) $\frac{2}{5}$ 16. $\frac{{}^{10}C_5 \times {}^{90}C_{15}}{{}^{100}C_{20}}$ 17. $\frac{3}{5}$

18. (i) $\frac{1}{8}$ (ii) $\frac{1}{8}$ (iii) $\frac{1}{56}$ 19. (i) $\frac{1}{35}$ (ii) $\frac{24}{35}$ (iii) $\frac{11}{35}$ 20. (a) $\frac{1}{10}$ (b) $\frac{1}{2}$

Website :- https://www.arjun00.com.np