1

DATA CLASSIFICATION AND TABULATION

Unit Structure

1.0 INTRODUCTION

After collecting, the desired data the first step to be taken is to classify and tabulate the data. Unwidely, unorganized mass of collected data is not capable of being easily associated or interpreted. Unorganized data are not fit for further analysis and interpretation. In order to make the data simple and easily understandable, simplify them in such a way that irrelevant data are removed and their significant feature s are stand out prominently. The procedure adopted for this purpose is known as classification and tabulation. The classification and tabulation provide a clear picture of the collected data and on that basis the further processing is decided.

1.1 OBJECTIVES OF DATA CLASSIFICATION

- To consolidate the volume of data so that the similarities and differences can be easily understood.
- To facilitate comparison and highlights the significant characteristic of data

- To eliminate unnecessary details
- To allow one to get a mental picture of the information and helps in drawing inferences.
- To allow a statistical treatment of the data collected

1.2 MEANING OF CLASSIFICATION

Classification is the grouping of related facts into classes.

"The process of arranging things in groups or classes according to their resemblances and affinities and gives expression to the unity of attributes that may subsist amongst a diversity of individuals". - Connor [1997)

"Classification is the process of arranging data into sequences and groups according to their common characteristics or separating them into different but related parts". – Secristi

"The process of grouping large number of individual facts and observations on the basis of similarity among the items is called classification". - Stockton & Clark

Usually the data can be collected through questionnaire, schedules or response sheets. This collected data need to be consolidated for the purpose of analysis and interpretation. This process is known as Classification and Tabulation. We can include a huge volume of data in a simple statistical table and one can get an outline about the model by observing the statistical table rather the raw data. To construct diagrams and graphs, it is essential to tabulate the data.

For Example, letters in the post office are classified according to their destinations viz., Delhi, Madurai, Bangalore, Mumbai etc.,

1.3 REQUISITES OF IDEAL CLASSIFICATION

It should be unambiguous: There should be no uncertainty or ambiguity. Classes should be defined rigidly, so as to avoid any ambiguity.

It should be flexible: The classification should be enough to accommodate change, amendment and inclusion in various classes in accordance with new situations.

It should be homogeneous: Units of each class should be homogeneous. All the units included in a class or group should be present according to the property on basis of which the classification was done.

It should be suitable for the purpose: The composition of the class should be according to the purpose.

For example: To find out the economic condition of the persons, create classes on the basis of income.

It should be stable: Stability is necessary to make data comparable and to make out meaningful comparison of the results. This means that the classification of data set into different classes must be performed ia a way, that whenever an investigation is carried out, there is no change in classes and so the results of the investigation can be compared easily.

It should be exhaustive: Each and every item of data must belong to a particular class. An ideal classification is one that is free from any residual classes such as others or miscellaneous, as they do not state the characteristics clearly and completely.

It should be mutually exclusive: The classes should be mutually exclusive.

1.4 TYPES OF CLASSIFICATION

The data can be classified on the basis of following four criteria:

a. Geographical Classification

b. Chronological Classification

c. Qualitative Classification

d. Quantitative Classification

a. Geographical Classification: When data are classified with reference to geographical locations such as countries, states, cities, districts etc, it is known as geographical classification.

For Example: The production of rice in different states of India, production of wheat in different countries etc.

State-wise comparison of Rice production in India

Name of State	Production of Rice (Metric tonnes/hectares)
Andra Pradesh	7.49
Bihar	6.5
Chhattisgarh	6.09
Punjab	11.82
Tamil Nadu	7.98
Utter Pradesh	12.5
West Bengal	15.75

Geographical Classification are usually listed in alphabetical order for easy reference. Items may also be listed by size to emphasis the important areas as in ranking the States by population.

b. Chronological Classification: Classification where data are grouped according to time is known as chronological classification

For Example the population of India from 1931 to 2001.

Population of India from 1931 to 2001

Year	Population (in millions)
1931	276
1941	313
1951	357
1961	438
1971	536
1981	634
1991	846
2001	1002

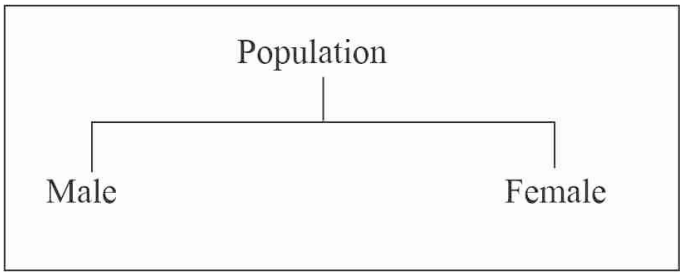
Time series are usually listed in chronological order, normally starting with the earliest period. When the major emphasis falls on the most recent events, a reverse time order may be used.

c. Qualitative Classification: In Qualitative classification, data are classified on the basis of some attributes or quality such as gender, religion, literacy, marital status etc. In this type of classification, the attribute under study cannot be measured. It can only be found out whether it is present or absent in the units of study.

For example, if the population to be classified in respect to one attribute, say gender, then we can classify them into two classes namely males and females.

Thus when only one attribute is studied two classes do formed, one possess the attribute and the other not possessing the attribute. This type of classification is known as simple or dichotomous classification.

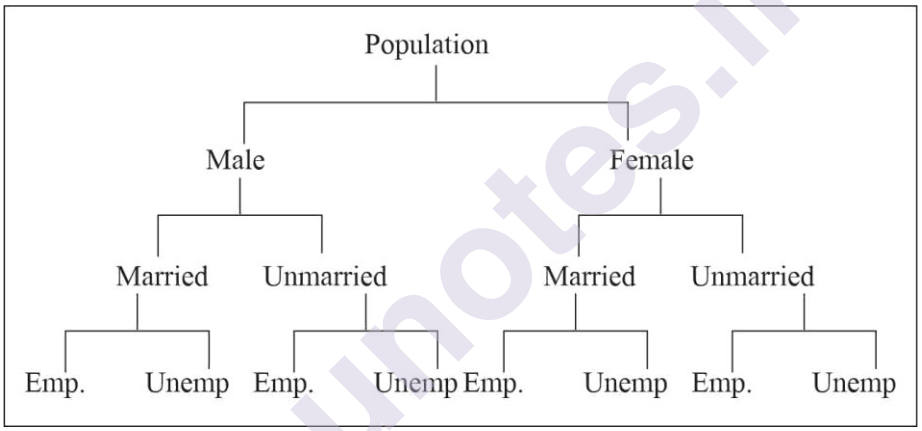
For example, if the population under study is divided into categories as follows:



If instead of forming only two classes we further divide the data on the basis of some attribute or attributes so as to form several classes, the classification is known as manifold classification.

For example, we may first divide the population into males and females on the basis of the attribute gender, each of these classes may subdivided into married and unmarried on the basis of marital status. Further classification can be made on the basis of attribute say, employment.

Example of manifold classification is as follows:



d. Quantitative Classification: It refers to the classification of data according to some characteristics that can be measured such as height, weight, profits, income, sales etc.

For example, the students of the school may be classified according to weight as follows:

Weight (in kg)	No. of Students
40-50	40
50-60	160
60-70	110
70-80	200
80-90	90
Total	**600**

Such a distribution is known as empirical frequency distribution or simple frequency distribution.

In this type of classification, there are two elements,

a. The variable (weight in above example)

b. The frequency (the number of students in each class) i.e. there are 40 students having weight ranging from 40 to 50 kg, 160 students having weight 50 to 60 kg and so on.

Thus, we can find out the ways in which the frequencies are distributed.

1.5 FREQUENCY DISTRIBUTION

A frequency distribution refers to data classified on the basis of some variable that can be measured such as prices, wages, age, height, weight. The term variable refers to the characteristic that varies in amount or magnitude in a frequency distribution.

A variable may be either **discrete or continuous**.

A discrete variable is variable whose value is obtained by counting. A discrete variable is that which can vary only by finite jumps.

For example, number of children, number of students in a class.

A continuous variable, also called continuous random variable is a variable whose value is obtained by measuring. In a continuous variable, data are obtained by numerical measurements rather than counting.

For example, when a student grows from 40 kg to 50 kg., his weight passes through all values between these limits.

Following are the examples of discrete and continuous frequency distribution.

Discrete Frequency Distribution:

No. of children	No. of families
0	20
1	50
2	90
3	70
4	10

Continuous Frequency Distribution:

Weight (in kg)	No. of Students
40-50	40
50-60	160
60-70	110
70-80	200
80-90	90
Total	**600**

1.5.1 Formation of a Discrete Frequency Distribution

The formation of discrete frequency distribution is quite simple. The number of times a particular value is repeated is noted down and mentioned against that values instead of writing that value repeatedly. In order to facilitate counting prepare a column of tallies. In another column, place all possible values of variable from lowest to the highest. Then put a bar (vertical line) opposite the particular value to which it relates. To facilitate counting, blocks of five bars are prepared and some space is left in between each block. Finally count the number of bars and get the frequency.

Example 1: The daily wages in Rs. paid to the workers are given below. Form the Discrete Frequency Distribution.

300, 200, 200, 100, 100, 200, 100, 100, 400, 100, 100, 200, 400, 300, 100, 300, 100, 100, 200, 100, 100, 100, 200, 100, 200, 100, 100, 100, 400, 200.

Solution: Frequency distribution of daily wages in Rs

Daily wages (in Rs.)	Tally Marks	No. of workers
100	~~IIII IIII IIII I~~	16
200	~~IIII~~ III	08
300	III	03
400	III	03
		Total = 30

1.5.2 Formation of a Continuous Frequency Distribution

The following technical terms are important when a continuous frequency distribution is formed.

a. Class Limits: Class limits are the lowest and the highest values that can be included in a class. The two boundaries of class are known as the lower limit and the upper limit of the class. The lower limit of the a class is the value below which there can be no item in the class. The upper limit of the a class is the value above which there can be no item in the class.

For example, for the class 20 – 40, 20 is the lower limit and 40 is the upper limit. If there was an observation 40.5, it would not be included in this class. Again if there was an observation of 19.5, it would not be included in this class.

b. Class Intervals: The difference between upper and lower limit of the class is known as class interval of that class.

For example, in the class 20 – 40, the class interval is 20 (i. e. 40 minus 20). An important decision while constructing a frequency distribution is about the width of the class interval i. e. whether it should be 10, 20, 50, 100, 500 etc. It depends upon the range in the data, i. e. the differencebetween the smallest and largest item, the details required and number of classes to be formed, etc.

Following is the simple formula to obtain the estimateof appropriate class interval,

$i = \frac{L-S}{k}$,

where i = Class interval,

L = largest item

S = smallest item

K = the number of classes.

For example, if the salary of 100 employees in a company undertaking varied between Rs. 1000 to 6000 and we want to form 10 classes then the class interval would be

$i = \frac{L-S}{k}$

L = 6000, S = 1000, k = 10

$i = \frac{L-S}{k}$

$= \frac{6000-1000}{10}$

= 500

The staring class would be 1000 – 1500, 1500 – 2000 and so on.

The question now is how to fix the number of classes i.e. k. The number can be either fixed arbitrarily keeping in view the nature of problem under study or it can be decided with the help of Sturge's Rule.

According to Sturge's Rule, number of classes can be determined by the formula:

$k = 1 + 3.322 \log N$

where N = total number of observations and

log = logarithms of the number.

Therefore, if 10 observations are there, the number of classes shall be,

$k = 1 + (3.322 \times 1) = 4.322$ or 4 [$\because \log 10 = 1$]

Therefore, if 100 observations are there, the number of classes shall be,

$k = 1 + (3.322 \times 2) = 7.644$ or 8 [$\because \log 100 = 2$]

It should be noted that since log is used in the formula, the number of classes shall be between 4 and 20. It cannot be less than 4 even if N is less than 10 and if N is 10 lakh, k will be $1 + (3.322 \times 6) = 20.9$ or 21.

c. Class Frequency: The number of observations corresponding to a particular class is known as the frequency of that class or the class frequency.

d. Class Mid-point or Class mark: Mid point of a class is calculated for further calculations in statistical work.

Mid-point of a class = $\frac{Upper\ limit\ of\ the\ class + Lower\ limit\ of\ the\ class}{2}$

1.6 METHODS OF DATA CLASSIFICATION

There are two methods of classifying the data according to class intervals.

a. Exclusive method: When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class it is known as the exclusive method of classification.

Weight (in kg)	No. of Students
40-50	40
50-60	160
60-70	110
70-80	200
80-90	90

In the above example, there are 40 students whose weight is between 40 to 49.99 kg. A student whose weight is 50 kg is included in the class 50 – 60.

b. **Inclusive method:** In this method, the upper limit of one class is included in that class itself.

Weight (in kg)	No. of Students
40-49	40
50-59	160
60-69	110
70-79	200
80-89	90

In the class 40 – 49, we include students whose weight is between 40 kg and 49 kg. If the weight of the student is exactly 50 kg he is included in next class.

Example 2: Prepare a frequency distribution for the students marks data.

25	85	41	70	85	55	85	55	72	72	50	90
52	68	72	52	91	53	79	75	60	35	65	80
70	70	36	66	55	80	72	41	88	60	45	78
42	90	66	47	80	88	91	82	50	52	55	72
68	65										

Solution: Since the lowest value is 25 and the largest value is 91, we take class intervals of 10.

Marks	Tally Marks	Frequency
25 – 35	I	01
35 – 45	~~IIII~~	05
45 – 55	~~IIII~~ III	08
55 – 65	~~IIII~~ I	06
65 – 75	~~IIII IIII~~ IIII	14
75 – 85	~~IIII~~ II	07
85 - 95	~~IIII~~ IIII	09
		Total = 50

Example 3: Prepare a frequency distribution for the following data by taking class interval such that their mid values are 17, 22, 27, 32 and so on.

30	30	36	33	42	27	22	41	30	42	30	21
54	36	31	40	28	19	48	26	48	15	37	16
17	54	42	51	44	32	42	31	21	25	36	22
41	40	46									

Solution: Since we have to classify the data in a such manner that the mid values are 17, 22, 27, 32 and so on The first class interval should be 15 – 19 (mid-value = (15 + 19)/2 = 17).

Variable	Tally Marks	Frequency
15 – 19	IIII	4
20 – 24	IIII	4
25 – 29	IIII	4
30 – 34	~~IIII~~ III	8
35 – 39	IIII	4
40 – 44	~~IIII~~ IIII	9
45 – 49	III	3
50 - 54	III	3
		Total = 39

1.7 BI-VARIATE FREQUENCY DISTRIBUTION

We know that if frequency distributions are involving one variable then it is called as univariate frequency distribution. In many situations simultaneous study of two variables becomes necessary. The data so classified on the basis of two variables give rise to what is called a bivariate frequency distribution. While preparing a bivariate frequency distribution, the same considerations of classification apply as for univariate frequency distribution i. e. the values of each variable. If the data corresponding to one variable, say X, is grouped into m classes and the data corresponding to the other variable, say Y, is grouped into n classes then the bivariate table will consist of m x n cells. By going through the different pairs of the values (X, Y) of the variables and using tally marks we can find the frequency of each cell and thus form bivariate frequency distribution.

Example 4: Construct a bivariate frequency distribution table of the marks obtained by the students in Mathematics (X) and Physics (Y)

Marks in Mathematics (X)	Marks in Physics (Y)
37	30
20	32
46	41
28	33
35	29
26	43
41	30
48	21
32	44
23	38
20	47
39	24
47	32
33	21
27	20
26	21

Solution: Let X: Marks in Mathematics and Y: Marks in Physics

X → Y	20 – 30	30 – 40	40 - 50	Total
20 – 30	II	III	I	6
30 – 40	III	I	II	6
40 - 50	II	I	I	4
Total	7	5	4	16

Two way frequency table showing marks in Mathematics and marks in Physics.

1.8 TABULATION OF DATA

The simplest and most revealing devices for summarizing data and presenting them in a meaningful manner is the statistical table. After classifying the statistical data, next step is to present them in the form of tables. A table is a systematic organization of statistical data in rows and columns. The purpose of a table is to simplify the presentation and to facilitate comparisons. The main objective of tabulation is to answer various queries concerning the investigation. Tables are very helpful for doing analysis and drawing inferences from them.

Classification and tabulation go together, classification being the first step in tabulation. Before the data are put in tabular form, they have to be classified.

1.9 Objectives of Tabulation

- **To simplify complex data:** It reduces raw data in a simplified and meaningful form. The reader gets a very clear idea of what the table present. It can be easily interpreted by a common person in less time.
- **To facilitate comparison:** Since the table is divided into rows and columns, for each row and column there is total and subtotal, the relationship between different parts of data can be done easily.
- **To bring out essential features of data**: It brings out main features of data. It presents facts clearly and precisely without textual explanation.
- **To give identity to the data**: when the data are arranged in a table with title and number, they can be differently identified.
- **To save space**: Table saves space without sacrificing the quality and quantity of data.

1.10 PARTS OF TABLE

Generally, a table should be comprised of the following components:

1. **Table Number:** Each table must be given a number. Table number helps in distinguishing one table from other tables. Usually tables are numbered according to the order of their appearance in a chapter. For example, the first table in the first chapter of a book should be given number 1.1 and second table of the same chapter be given 1.2. Table number should be given at its top or towards the left of the table.
2. **Title of the Table:** The title is a description of the contents of the table. Every table must be given suitable title. A complete title has to answer the questions what categories of statistical data are shown, where the data occurred and when the data occurred. The title should be clear and brief. It is placed either just below the table number or at its right.

3. **Caption:** Caption refers to the column headings. It may consist of one or more column headings. Under one column there may be sub heads. The caption should be clearly defined and placed at the middle of the column. If the different columns have different units, the units should be mentioned with the captions.

4. **Stub:** Stub refers to the rows or row heading. They are at extreme left of the table. The stubs are usually wide than column headings but they are as narrow as possible.

5. **Body:** It is most important part of the table. It contains number of cells. Cells are formed by intersection of rows and columns. The body of the table contains numerical information.

6. **Headnote:** It is used to explain certain points relating to the whole table that have not included in the title, in the caption or stubs. It is placed below the title or at the right hand corner of the table. For example, the unit of measurement is frequently written as a headnote, such as "in thousands", "in crores", etc.

7. **Footnotes:** It helps in clarifying the point which is not clear from the title, captions or stubs. It is placed at the bottom of a table.

There are different ways of identifying the footnotes. One is numbering them consecutively with small numbers 1, 2, 3 or letters a, b, c, d. Another way identifies the first footnote with one star (*), second footnote with two stars (**), third footnote with three stars (***) and so on. Sometimes instead of * , +,@,£ etc used.

Format of a Table

Table Number

Title **Headnote**

Stub heading	Caption heading – Column Heading
Stub entries	Body

Footnote

1.11 TYPES OF TABLES

There are three basis of classifying table.

- Purpose of table
- Originality of a table
- Construction of a table

1. **Tables according to Purpose:** According to purpose, there are two kinds of tables.

   i. **General Purpose Table:** It provide information for general use or reference. General Purpose Tables also known as reference tables or repository tables. They usually contains detailed information and are not constructed for specific discussion. These tables are generally attached to some official reports like Census Reports of India.

   ii. **Special Purpose Table:** Special purpose table is that table which is prepared with some specific purpose. These are the small tables limited to the problem under consideration. It is also known as **Summary Tables**. When attached to a report they are found in the body of the text.

2. **Tables according to Originality:** On the basis of originality, tables are of two types.

   i. **Original Table:** An original table is that in which data are presented in the same form and manner in which they are collected.

   ii. **Derived Table:** A derived table is that in which data are not presented in the same form and manner in which they are collected but the data are first converted into ratios or percentage and then presented.

3. **Tables according to Construction:** According to construction, tables are of two kinds:

   i. **Simple or one way Table:** In this table only one characteristics of data is shown.

   For example: Following table shows number students in a college**.**

   Number of students in a college

Class	Number of students
MCA I	240
MCA II	220
Total	**460**

   ii. Complex Table: A complex table is one which shows more than one characteristic of the data. On the basis of characteristics shown, these tables may be further classified as:

      a. Double or Two way Table: A two way table is that shows two characteristics of the data.

For example: Following table shows number students in a college according to gender.

Number of students in a college

(According to class and gender)

Class	Number of students		Total
	Boys	Girls	
MCA I	180	60	240
MCA II	170	50	220
Total	**350**	**110**	**460**

b. Treble Table: A treble table is that shows three characteristics of the data.

For example: Following table shows number students in a college according to class, gender and location.

Number of students in a college

(According to class, gender and location)

Class	Number of students				Total
	Boys		Girls		
	Urban	Rural	Urban	Rural	
MCA I	80	100	40	20	240
MCA II	90	80	30	20	220
Total	**170**	**180**	**70**	**40**	**460**

c. Manifold Table: When more than three characteristics are shown in table, such table is called manifold table.

For example: Following table shows number students in a college according to class, gender, location and marital status.

Number of students in a college

(According to class, gender, location and marital status)

Class	Number of students								Total
	Boys				Girls				
	Urban		Rural		Urban		Rural		
	Married	Unmarried	Married	Unmarried	Married	Unmarried	Married	Unmarried	
MCA I	20	60	60	40	10	30	15	5	240
MCA II	40	50	50	30	10	20	15	5	220
Total	**60**	**110**	**110**	**70**	**20**	**50**	**30**	**10**	**460**

Example 5: Following information relates to the marks secured by 50 students. Present the information in the form of table.

Marks	0 – 10	10 – 20	20 – 30	30 - 40
Students	15	12	18	5

Solution:

Table showing marks of students

Marks	No. of students
0 – 10	15
10 – 20	12
20 – 30	18
30 – 40	5
Total	50

Example 6: Following information relates to the marks secured by 50 students. Present the information in the form of a two way table.

Marks	0 – 10	10 – 20	20 – 30	30 - 40
Boys	12	8	9	2
Girls	3	4	9	3

Solution:

Table showing marks of students

(According to gender)

Marks	No. of students		Total
	Boys	Girls	
0 – 10	12	3	15
10 – 20	8	4	12
20 – 30	9	9	18
30 – 40	2	3	5
Total	31	19	50

Example 7: Draft a blank table to show the distribution of information according to

a. Gender: Male and Female

b. Literacy: Literates and Illiterates

c. Age Group: 0 – 25, 25 - 50, 50 – 75 and 75 – 100

Solution:

Table showing the distribution of Population

(According to age, gender and literacy)

Age Group	Literates			Illiterates			Total		
	M	F	Total	M	F	Total	M	F	Total
0 - 25									
25 – 50									
50 – 75									
75 - 100									
Total									

1.12 SUMMARY:

After collection and editing of data the first step towards further processing is classification. Although the term "classification and tabulation" has been used, classification is the first step in tabulation. Items having common characteristics must be brought together before the data can be displayed in tabular form. Classification helps proper tabulation. Tabulation is a logical step of presenting statistical data after classification. Tabulation enables the data to be presented in a manner that suits for further statistical treatment and for making valid conclusions.

1.13 REFERENCES

Books: 1. Statistical Methods – S. P. Gupta

2. Business Statistics – Pearson

Websites:

https://www.brainkart.com/article/Bivariate-Frequency-Distributions_35069/

https://www.slideshare.net/bijayabnanda/ls-bs-3classification-tabulation-of-data

https://www.vedantu.com/commerce/tabulation

1.14 EXERCISE

Exercise 1: The number of children per family is given below. Form the Discrete Frequency Distribution.

1	7	8	7	9	0	2	4	6	7	2	3
5	5	9	3	4	12	3	4	4	0	6	3
5	5	2	3	7	4	6	5	2	8	3	4
5	6	7	2								

Exercise 2: Following data is relating to the daily number of car accidents during 30 days of month. Represent it in the Discrete Frequency Distribution.

3	4	4	5	3	4	3	5	7	6	4	4
3	4	5	5	5	5	5	3	5	6	4	5
4	4	6	5	6							

Exercise 3: The following is the number of female employees in different branches of commercial banks. Make a frequency distribution.

2, 4, 6, 1, 3, 5, 3, 7, 8, 6, 4, 7, 4, 4, 2, 1, 3, 6, 4, 2, 5, 7, 9, 1, 2, 10, 1, 8, 9, 2, 3, 1, 2, 3, 4, 4, 4, 6, 6, 5, 5, 4, 5, 8, 5, 4, 3, 3, 2, 5, 0, 5, 9, 9, 8, 10, 0, 4, 10, 10, 1, 1, 2, 2, 1, 8, 6, 9, 10

Exercise 4: Prepare a frequency distribution for following marks (out of 50) obtained in Mathematics by 60 students

21	10	30	22	33	5	37	12	25	42	15	39
26	32	18	27	28	19	29	35	31	24	36	18
20	38	22	44	16	24	10	27	39	28	49	29
32	23	31	21	34	22	23	36	24	36	33	47
48	50	39	20	7	16	36	45	47	30	22	17.

Exercise 5: Let us consider the following example regarding daily maximum temperatures in $^{\circ}C$ in a city for 50 days. Prepare a frequency distribution

28	28	31	29	35	33	28	31	34	29	25	27
29	33	30	31	32	26	26	21	21	20	22	24
28	30	34	33	35	29	23	21	20	19	19	18
19	17	20	19	18	18	19	27	17	18	20	21
18	19										

Exercise 6: Construct a bivariate frequency distribution table for the height and weight of 20 persons.

S. No.	Weight (in pounds)	Height (in inches)	S. No.	Weight (in pounds)	Height (in inches)
1	163	70	11	170	70
2	139	67	12	135	65
3	122	63	13	136	65
4	134	68	14	137	64
5	140	67	15	148	69
6	132	69	16	121	63
7	120	65	17	117	65
8	148	68	18	128	70
9	129	67	19	143	71
10	152	67	20	129	62

Exercise 7: Draft a blank table to show the distribution of information according to

a. Gender: Boys and Girls

b. Faculty: Science and Arts

c. Course: Management, Costing, Costing and finance

Exercise 8: Draft a blank table to show the distribution of information according to

a. Gender: Male and female

b. Salary Grade: 1000 – 5000, 5000 – 10000 and 10000 – 15000

c. Years: 2007 and 2008

Exercise 9: Present the following data of marks of 60 students in a form of a frequency table with 10 classes of equal width, one class being 40 - 49.

43	35	62	5	61	19	46	35	69	34	35	59
11	26	74	65	51	81	66	55	56	36	30	65
94	94	75	32	13	90	62	56	44	37	62	60
65	95	84	57	72	38	45	12	62	08	54	82
25	35	69	34	05	02	42	84	77	34	42	59

❄❄❄❄❄❄❄

2

DIAGRAMMATIC AND GRAPHIC PRESENTATION OF DATA

Unit Structure

2.0 INTRODUCTION

In the previous chapter, we have studied the importance and techniques of classification and tabulation that help to arrange the mass of collected data in a logical and summarize manner. However, it is a difficult and cumbersome task for common man and researcher to interpret the data. Too many figures are often confusing and may fail to convey the message effectively to those for whom it is meant.

To overcome this inconvenience, the most appealing way in which statistical results may be presented is through diagrams and graphs. A diagram is a visual form for presentation of statistical data, highlighting their basic facts and relationships. If we draw diagrams on the basis of the data collected they will be understood and appreciated by all. Everyday we

can find the presentation of stock market, cricket score etc. in news paper, television and magazines in the form of diagrams and graphs. In this chapter we will discuss some of the major types of diagrams, graphs and maps frequently used in presenting statistical data.

2.1 OBJECTIVES

After going through this chapter, students will able

- To explain the need and significance of diagrams and graphs.
- To describe various types of diagrams and explain how to present the data through an appropriate diagram
- To list out general rules for drawing diagrams
- To list out and differentiate between the major forms of diagrams and graphs
- To present frequency distribution in the form of various type of graphs

2.2 SIGNIFICANCE OF DIAGRAMS AND GRAPHS

Visual presentation of data means presentation of statistical data in the form of diagrams and graphs. Diagrams and graphs are extremely useful because of the following reasons

1. **They are attractive and impressive:** The data when presented in the form of diagrams and graphs, give a birds eye-view of the entire data and creates interest and leaves an impression on mind for a long period.
2. **They make data simple and intelligible:** Pictorial presentation helps in proper understanding of the data as it gives an interesting form of it.
3. **They make comparison easy:** Diagrams and graphs make comparison of data relating to different periods of time of different regions. Diagrams and graphs make quick and accurate comparison between two sets of data.
4. **They have universal applicability:** It is a universal practice to present the numerical data in the form of diagrams and graphs. In these days, it is an extensively used technique in the field of economics, business, education, health, agriculture etc.
5. **They give more information:** Graphs makes it possible to locate several measures of central tendency like Median, Mode, and Quartiles etc. They help in establishing trends of past performance and thus helps in forecasting.

6. **They save time and efforts:** Diagrams and graphs reduce the strain and save a lot of time in understanding the basic characteristics of the data.

7. **They have become an integral part of research:** Now a days it is difficult to find any research work without visual support. The reason is that this is the most convincing and appealing way of presenting the data.

2.3 ADVANTAGES AND LIMITATIONS OF DIAGRAMS (GRAPH)

Advantages:

1. The data presented in the form of diagrams is simplest and easy to understand.
2. Diagrams helps in making comparisons between two or more group or two or more periods.
3. Diagrammatic presentation makes the data more attractive and interesting.
4. They help to save time and efforts to understand the presentation.
5. Diagrams and graphs can be used as a source of reference for different researches and studies.

Limitations:

1. Diagrammatic presentation of data is just an approximation of the actual behavior of the variable
2. Only a limited set of data can be presented in the form of diagram.
3. Diagrams do not show small differences properly.
4. Diagrams can be used only for comparative study.
5. Diagrams can easily misused and misinterpreted.
6. It is not easy to come at final conclusion after seeing the diagram.

2.4 GENERAL RULES FOR DRAWING DIAGRAMS

1. **Title:** Title must be given to every diagram or graph. From the title one can know the idea contained in it. The title should be brief and self-explanatory. It is usually placed at the top or below it.

2. **Proper size and scale:**

   A diagram or graph should be of normal size and drawn with proper scale. The scale showing the values should be even numbers or in

multiples of 5 or 10 e.g. 10, 20, 30, 25, 50, 75. Odd values like 1, 3, 5 should be avoided. The scale in graphs specifies the size of the unit what it represents e.g. “million tons”, “number of persons in thousands” etc.

3. **Footnotes:** Footnote should be given at the bottom of the diagram to clarify certain points about the diagram.

4. **Index:** Every diagram or graph must be accompanied by an index. This illustrates different types of lines, shades or colors used in the diagram.

5. **Neatness and cleanliness:** A diagram should be neatly drawn and attractive.

6. **Simplicity:** Diagrams must be as simple as possible.

2.5 TYPES OF DIAGRAMS

In practice, large variety of diagrams is in use. Diagrams are classified on the basis of their length, width and shape. We will discuss the important types of diagrams which are more frequently used. For sake of application and simplicity several types of diagrams are categories under the following heads.

1. One dimensional diagrams

2. Two dimensional diagrams

3. Three dimensional diagrams

4. Pictogram or Ideographs

5. Cartograms or Statistical Maps

2.6 ONE DIMENSIONAL OR BAR DIAGRAMS

This is the most common type of diagrams. They are called one-dimensional diagrams because only length of the bar matters and not the width. For large number of observations lines may be drawn instead of bars to save space.

Merits of Bar diagrams:

1. They are easily understood.

2. They are simplest and easiest to make

3. They are simplest and easiest in comparing two or more diagrams.

Types of Bar Diagrams:

a. Simple bar diagram

b. Subdivided bar diagram

c. Multiple bar diagram

d. Percentage bar diagram

e. Deviation bars

2.6.1 Simple Bar Diagram: A simple bar diagram is used to represent only one variable. It should be kept in mind that, only length is taken into account and not width. Width should be uniform for all bars and the gap between each bar is normally identical. For example the figures of production. Sales, profits etc for various years can be shown by bar diagrams.

Example 1: Prepare a simple bar diagram for following data related to wheat exports.

Year	Exports (in million tons)
2001	177
2002	219
2003	420
2004	326
2005	202
2006	225

Solution:

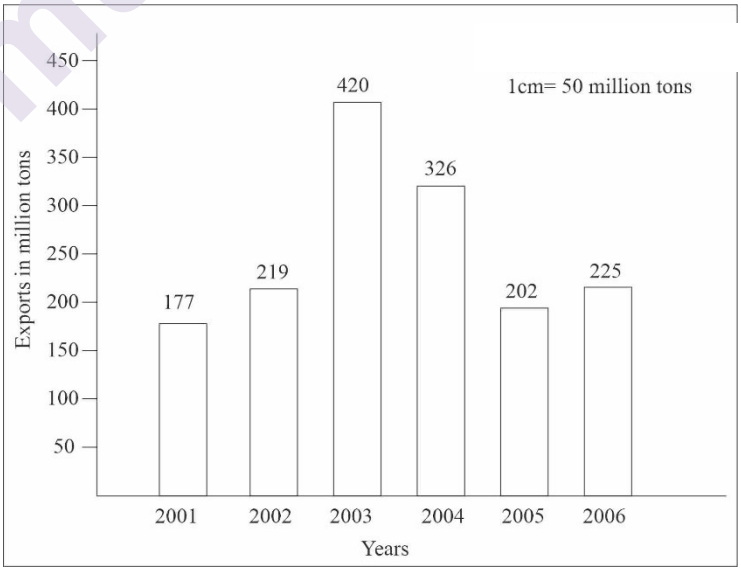


Figure 2.1 Simple Bar Diagram Showing the Wheat Exports in Different Years

2.6.2Subdivided Bar Diagram: In this diagram, one bar is constructed for total value of the different components of the same variable. Further, it is subdivide impropriation to the various components of that variable.

A bar is represented in the order of magnitude from the largest component at the base of the bar to the smallest at the end of the bar, but the order of various components in each bar is kept in the same order. Different shades or colors are used to distinguish between different components. To explain such differences, the index should be used in the bar diagram. The subdivided bar diagrams can be constructed both on horizontal and vertical bases.

Example 2: The following data shows the production of rice for the period 2010 to 2018. Represent the data by a subdivided bar diagram.

Year	Non-Basmati Rice (in Million metric tons)	Basmati Rice (in Million metric tons)	Total (in Million metric tons)
2010	29	35	64
2011	35	33	68
2012	25	35	60
2013	40	30	70
2014	42	32	74
2015	32	40	72

Solution:

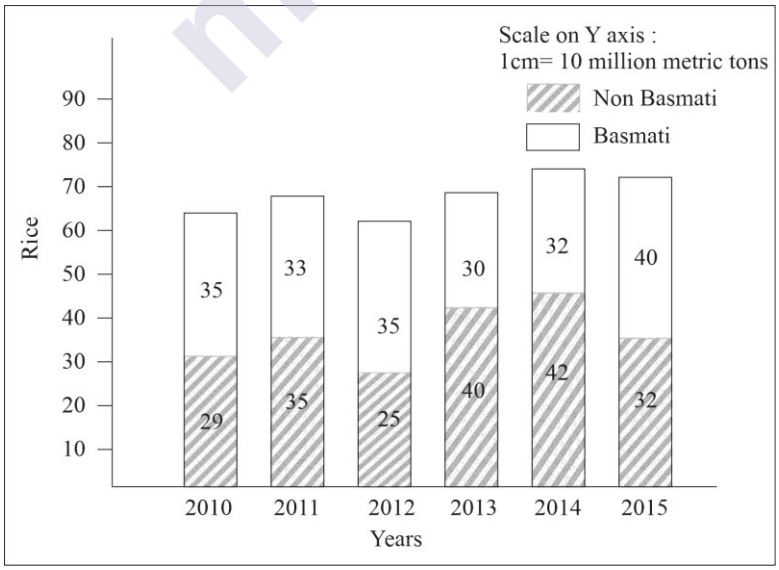


Figure 2.2 Subdivided Bar Diagram Showing the production of Rice (in Different Years)

2.6.3 Multiple Bar Diagram: Whenever the comparison between two or more related variables is to be made, multiple bar diagram should be preferred. In multiple bar diagrams two or more groups of interrelated data are presented. The technique of drawing such type of diagrams is the same as that of simple bar diagram. The only difference is that since more than one components are represented in each group, so different shades, colors, dots or crossing are used to distinguish between the bars of the same group.

Example 3: Represent the following data by a multiple bar diagram.

Class	Physics	Chemistry	Mathematics
Student A	50	63	57
Student B	55	60	68
Student C	48	60	55

Solution:

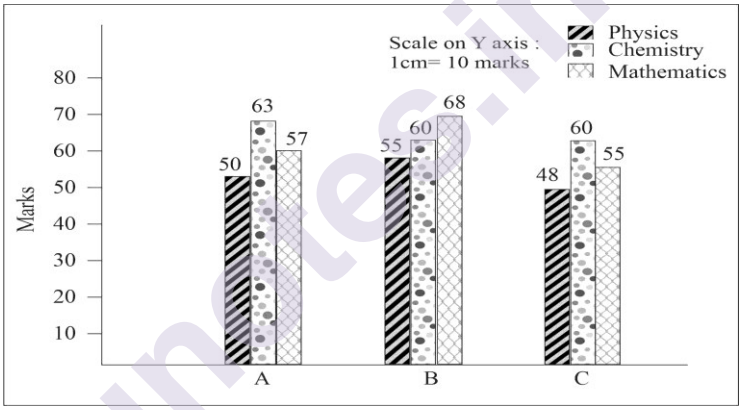


Figure 2.3 Multiple Bar Diagram

2.6.4 Percentage Bar Diagram: Percentage bars are particularly useful in statistical work which requires the representation of the relative changes in data. When such diagrams are prepared, the length of the bars is kept equal to 100 and segments are cut in these bars to represent the percentages of an average.

Example 4: Draw percentage bar diagram for following data.

Particulars	Cost Per Unit (2010)	Cost Per Unit (2020)
Material	22	35
Lobour	30	40
Delivery	10	20
Total	62	95

Solution: Express the values in terms of percentage for both the years.

Particulars	Cost Per Unit (2010)	% Cost	Cumulative % cost	Cost Per Unit (2020)	% Cost	Cumulative % cost
Material	22	35.48	35.48	35	36.84	36.84
Lobour	30	48.38	83.86	40	42.10	78.94
Delivery	10	16.12	99.98	20	21.05	99.99
Total	62	100		95	100	

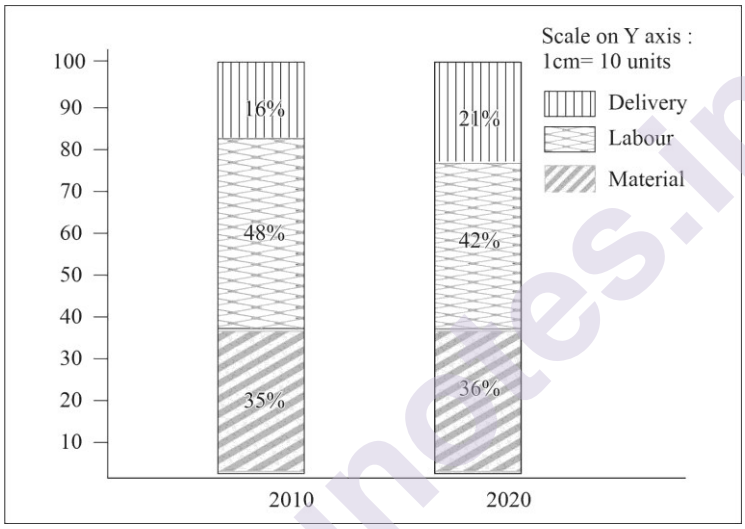


Figure 2.4 Percentage Bar Diagram

2.6.5Deviation Bar Diagram: Deviation bars are used for representing net quantities – excess or deficit. i. e. net profit, net loss, net exports or imports etc For representing net quantities excess or deficit, i.e. net profit, net loss, net exports, net imports, etc., This kind of bars represent both positive and negative values. The values which are positive can be drawn above the base line and negative values can be drawn below it.

Example 5: Draw deviation bar diagram for following data.

Year	Sales	Profits
2010	24%	29%
2011	15%	-10%
2012	23%	-5%

Solution:

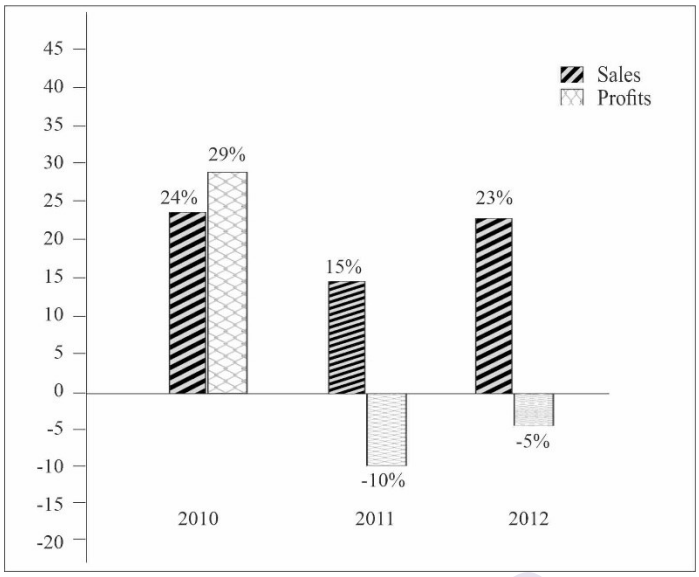


Figure 2.5 Deviation Bar Diagram for sales and profits

2.7 TWO DIMENSIONAL DIAGRAMS

In one-dimensional diagrams, only length of the bar is considered and comparison of bars are done on the basis of length only. In two-dimensional diagrams the length as well as width of the bars is considered. Thus, the area of the bars represent the given data. Two-dimensional diagrams are also known as surface diagrams or area diagrams.

Types of Two Dimensional Diagrams:

a. Rectangles
b. Squares
c. Circles

2.7.1 Rectangles: In rectangle diagram, given numerical figures are represented by areas of the rectangles. We know that area of rectangle = length x width. While constructing such a diagram both length and width are considered. We may represent the figures as they are given or may convert them to percentage and then subdivide the length into various components.

Example 6: Represent the following data of monthly expenditure (in rupees) of two families by suitable diagram.

Expenditure	Family A	Family B
Food	2400	1800
Clothing	1600	1200

Education	1000	800
Electricity	200	200
Miscellaneous	800	500

Solution: First convert the figures into percentage and take the cumulative sum of percentage.

Expenditure	**Family A**			**Family B**		
	Rs.	**%**	**Cumulative %**	**Rs.**	**%**	**Cumulative %**
Food	2400	40	40	1800	40	40
Clothing	1600	27	67	1200	27	67
Education	1000	17	84	800	18	85
Electricity	200	3	87	200	4	89
Miscellaneous	3200	13	100	500	11	100
Total	6000	100		4500	100	

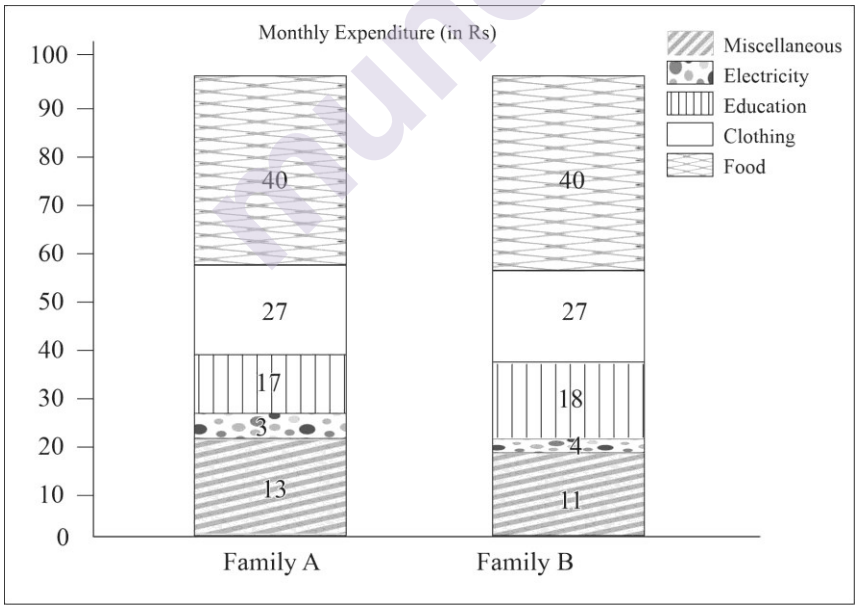


Figure 2.6 Diagram of Rectangles

2.7.2Squares: The rectangular method of diagrammatic presentation is difficult to use where the values of items vary widely. The method of drawing a square is simple. Take square roots of the given numerical

observations as sides of the corresponding squares and then select a suitable scale to draw the squares.

Example 7: Represent the following data of the number of hospitals in a city in 2000-05, 2005-10, 2010-15 and 2015-20 in square diagram.

Year	No. of Hospitals
2000-05	16
2005-10	64
2010-15	400
2015-20	576

Solution: Since there is a big gap between first year and last year, a square diagram is suitable here. To decide the side of a square consider following calculations.

Year	No. of Hospitals	Square Root	Side of square in cms = Square Root/4
2000-05	16	4	1
2005-10	64	8	2
2010-15	400	20	5
2015-20	576	24	6

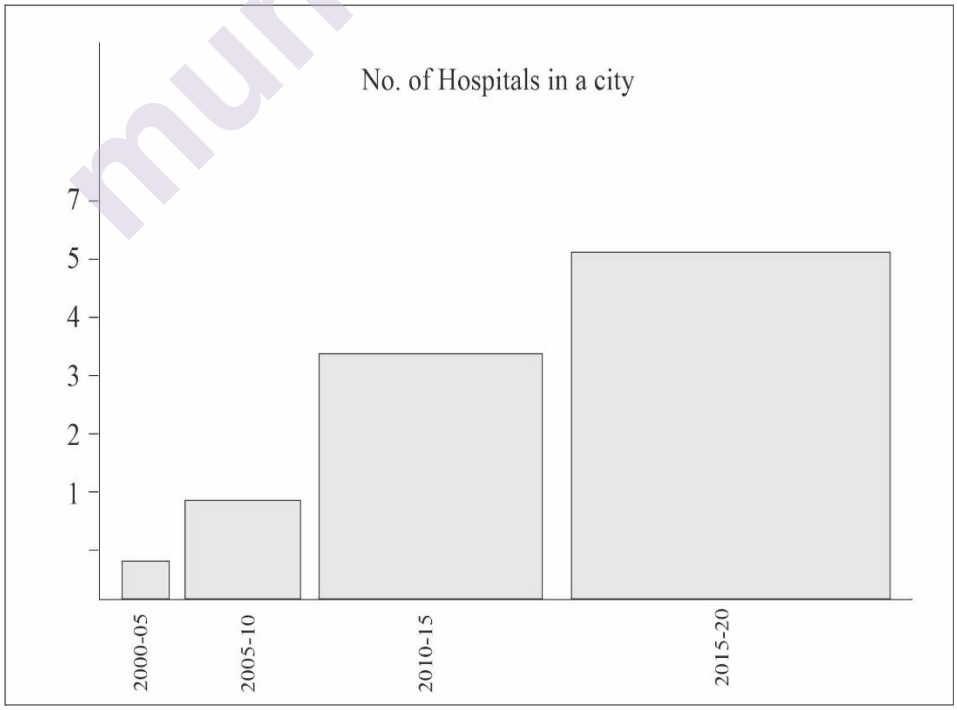


Fig 2.7 Diagram for Squares

2.7.3 Circles: As in square diagram, we took given figures/observations as the areas of the corresponding squares. Similarly, here we take given numerical figures/ observations as areas of the corresponding circles. The area of a circle is proportional to the squares of its radius. The radius of circles can be obtained by dividing the value of pie and taking square root. Circles can be used in all those cases in which squares are used.

Circles are difficult to compare and as such are not popular. When it is necessary to use circles, they should be compared on an area basis rather than on diameter basis.

Area of a circle = πr^2 where r is radius of circle

$\therefore \quad r^2 = \frac{Area}{\pi}$, $\pi = \frac{22}{7}$

$\therefore$ r = $\sqrt{\frac{Area}{\pi}}$

Example 8: Represent the Example 7 with the help of circles.

Year	No. of hospitals
2000-05	16
2005-10	64
2010-15	400
2015-20	576

Solution:

Year (I)	No. of Hospitals (n) (II)	n / ($\frac{22}{7}$) (III)	Square Root of [n / ($\frac{22}{7}$)] (IV)	Col (IV)/2 (V)
2000-05	16	5.09	2.25	1.125
2005-10	64	20.36	4.51	2.255
2010-15	400	12	11.28	5.64
2015-20	576	183.27	13.54	6.77

[Note: To get smaller value of radius of the circle divide each figure in Col (IV) by 2]

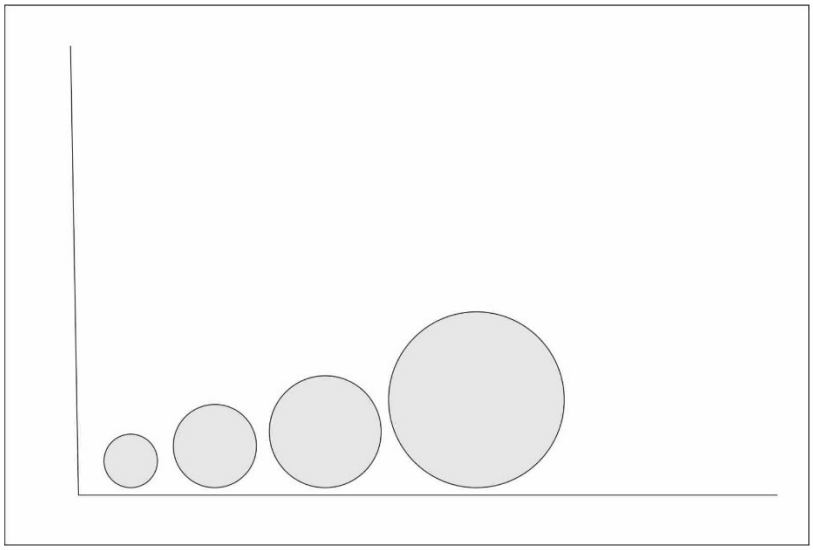


Fig 2.8 for Circles

Pie Diagram

A pie Diagram is a type of graph that displays data in a circular graph. The pieces of the graph are proportional to the fraction of the whole in each category. Pie diagrams are very popularly used in practice to show percentage breakdown. While making comparisons, pie diagrams should be used on a percentage basis and not on absolute basis.

How to create pie diagrams:

1. Take a total of all observations
2. Divide each observation by total and multiply by 100 to get percent. (if instead of percentage, observations are given)
3. Next to know how many degrees for each “pie sector” we need, we will take a full circle of 360° and follow the calculations as below:
4. The central angle of each component = (Value of each component/sum of values of all the components) x 360°
5. Draw a circle of appropriate size with compass and use the protractor to measure the degree of each sector.

In laying out the sectors for pie diagram, it is common practice to begin the largest component sector of pie diagram at 12 O’clock position on the circle. The other component sectors are placed in clockwise direction in descending order of magnitude. Give descriptive label for identification of each sector.

Example 9: Draw the pie diagram for the following data of cost of construction of house.

Bricks	15%
Steel	35%
Cement	20%
Labour	20%
Supervision	10%

Solution: Here values are given in percentage.

The angle at the Centre is given by

$\frac{Percentage\ outlay}{100}$ X 360 = Percentage outlay X 3.6

Sector	**Percentage**	**Angle outlay**
Bricks	15	15 x 3.6 = 54
Steel	35	35 x 3.6 =126
Cement	20	20 x 3.6 =72
Labour	20	20 x 3.6 =72
Supervision	10	10 x 3.6 =36
Total	100	360

(Note: The angles have been arranged in ascending order)

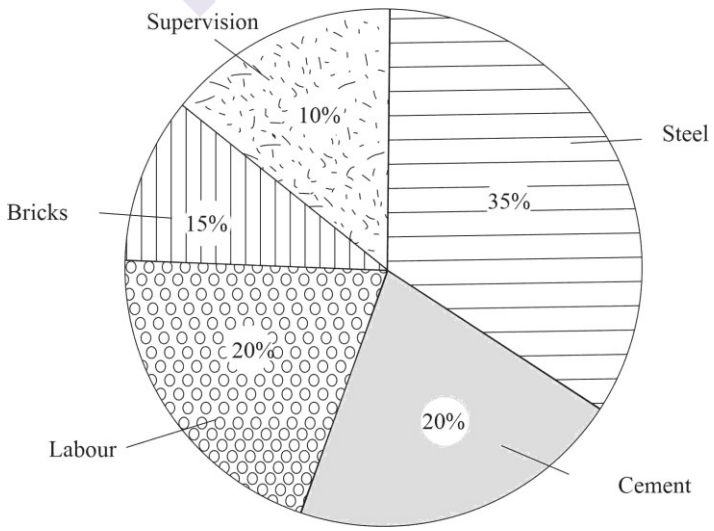


Fig 2.9 Pie Diagram

2.8 THREE DIMENSIONAL DIAGRAMS

In three dimensional diagrams three things namely, length, width and height have to be considered. Three-dimensional diagrams, also known as volume diagrams consist of cubes, cylinders, spheres etc. Such diagrams are used where the range of difference between the smallest and the largest value is very large. For example, if two values are in the ratio of 1:1000 and if bar diagram are used to represent them, the shortest bar would be of one-thousandth part of the largest bar. If squares or circles are used then the side of the square or the radius of one circle would be proportionately too large or too small than the other. If cubes are used then their sides would be in the ratio of 1:10. This example makes it clear that three-dimensional diagrams have an important role to play when the gap between the smallest and the largest value is very large.

The disadvantage of three-dimensional data is the side of a cube must be proportionate to the cube root of the magnitude to be represented. It is very difficult for the eye to read precisely such diagrams and hence they are not recommended for statistical presentation.

2.9 PICTOGRAM OR IDEOGRAPHS

A pictogram is a chart or graph which uses pictures to represent data in a simple way. They are very popularly used in presenting statistical data. They are set out the same way as a bar chart but used pictures instead of bars. Pictures are attractive and easy to comprehend. When pictograms are used, data are represented through a pictorial symbol that is carefully selected.

The pictorial symbol should be self-explanatory. For telling story about computer, the symbol of computer must be used.

Following points to be considered while selecting a pictorial symbol:

1. A symbol must be represent a general concept like men, women, car, not an individual.
2. A symbol should be clear and interesting.
3. A symbol must be clearly different from every other symbol.
4. A symbol should suit the size of paper not too small or too big

Changes in figures/numbers are shown by more or fewer symbols, not by larger or smaller ones.

Example 10: The following table shows the number of cars sold by a company for the months January to March. Construct a pictograph for the table.

Month	Number of Cars
January	120
February	200
March	260
April	100
May	180
June	220

Solution:

January	
February	
March	
April	
May	
June	

Advantages of a Pictograph:

1. Express a large amount of information or data in a simple form.
2. Since they make the use of symbols, pictographs attract attention, i.e. it is an attractive way to represent data.
3. Pictographs are easy to read since all the information is available at one glance.
4. Facts portrayed in pictorial form are generally remembered longer than facts presented in tables.

Disadvantage: Pictographs give only an overall picture; they do not give minute details.

2.10 CARTOGRAMS OR STATISTICAL MAPS

Cartograms or statistical maps are used to give quantitative information on a geographical basis. They are thus used to represent spatial distributions. A cartogram refers to a map through which statistical information are represented in different ways such as shades, dots, pictograms, columns. A cartogram is a type of graphic that depicts attributes of geographic objects as the object's area.

There are three main types of cartograms, each have a very different way of showing attributes of geographic objects.

1. Non-contiguous: It is the simplest and easiest type cartogram to make. In this, the geographic objects do not have to maintain connectivity with their adjacent objects. This connectivity is called topology.
2. Continuous: In this type, the objects remain connected with each other but due to this, there is distortion in shape.
3. Dorling cartograms: This is very effective cartogram method but it maintains neither shape, topology nor object centroids. To create a Dorling cartogram, instead of enlarging or shrinking the objects themselves, the creator will replace the objects with a uniform shape, normally a circle, of the appropriate size.

Statistical maps should be used only where geographic comparisons are of primary importance and where approximate measures will suit. For more accurate representation of size, bar charts are preferable. Maps are sometimes combined, which are drawn in the appropriate areas.

2.11 EXPLORATORY DATA ANALYSIS

Exploratory data analysis is a powerful way to explore a data set. Exploratory data analysis (EDA) is used to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. EDA can be used for data cleaning, for subgroup analyses or simply for understanding your data better. An important initial step in any data analysis is to plot the data.

We use EDA for following reasons

1. Detection of mistakes
2. Checking assumptions
3. Preliminary selection of appropriate models
4. Determining relationship among the explanatory variable

Types of Exploratory Data Analysis:

1. Univariate Non Graphical
2. Multivariate Non Graphical
3. Univariate Graphical
4. Multivariate Graphical

1. **Univariate Non Graphical:** This is simplest form of EDA. In this we use only one variable. The goal of this is to know the sample distribution and make observations about the population. The characteristics of population distribution include measures of central tendency, standard deviation, variance, skewness and kurtosis.
2. **Multivariate Non-Graphical:** Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross tabulation. For two variables, cross tabulation is performed by making a two-way table with column headings that match the levels of one-variable and row headings that match the levels of the other variable. Then filling in the counts of all subjects that share a pair of levels.
3. **Univariate Graphical:** Non-graphical are methods are quantitative and objective; they do not give a full picture of the data. Graphical methods are qualitative and involve a degree of subjective analysis. Common types of univariate graphics are Histogram, stem and leaf plot, Boxplots etc.
4. **Multivariate Graphical:** Multivariate graphical data uses graphics to display relationship between two or more sets of dat. Common types of multivariate graphics are Scatterplot, Run chart, Heat map, Multivariate chart and Bubble chart.

One should always perform appropriate EDA before further analysis of data. Perform whatever steps are necessary to become more knowledgeable in your data, check for common mistakes, learn about variable distributions and study about relationships between variables.

2.12 STEM AND LEAF DISPLAYS

A stem and leaf plot is a unique table where values of data are split into a stem and leaf. The first digit or digits will be written in stem and the last digit will be written in leaf. A stem and leaf display is used for presenting quantitative data in a graphical format. It is similar to histogram with the difference that in histogram, bars are used to compare data and in case of stem leaf plot, leaves represents actual numbers to be compared.

To construct a stem leaf display, sort the observations in ascending order. Then determine what the stems will represent and what the leaves will represent. The leaf contains the last digit of the number and the stem contains all of the other digits. The stem and display is drawn with two columns separated by a vertical line. The stems are listed to the left of the vertical line. It is important that each stem is listed only once and that no number is skipped, even if it means that some stems have no leaves. The leaves are listed in increasing order in a row to the right of each stem.

It gives the quick overview of the distribution. They retain most of the raw numerical data. Stem and leaf displays are useful for displaying the relative density and shape of the data.

Example 11: Prepare leaf stem plot for the following data.

31, 49, 19, 62, 50, 24, 45, 23, 51, 32, 48, 55, 60, 40, 35, 54, 26, 57, 37, 43, 65, 50, 55, 18, 53, 41, 50, 34, 67, 56, 44, 54, 57, 39, 52, 45, 35, 51, 63, 42

Solution: Sort the data in ascending order.

18, 19, 23, 24, 26, 31, 32, 34, 35, 35, 37, 39, 40, 41, 42, 43, 44, 45, 45, 48, 49, 50, 50, 50, 51, 51, 52, 53, 54, 54, 55, 55, 56, 57, 57, 60, 62, 63, 65, 67

Choose step as largest place value. Here it is 10. So each step will represent 10 units.

Group the numbers as per stem value.

18	19												
23	24	26											
31	32	34	35	35	37	39							
40	41	42	43	44	45	45	48	49					
50	50	50	51	51	52	53	54	54	55	55	56	57	57
60	62	63	65	67									

Draw the stem numbers at 10's place and leaves at 1's place digit.

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

2.13 SUMMARY

Statistical data not only requires a careful analysis but also ensures an attractive and communicative display. In order to achieve this objective, we discussed the techniques of diagrammatic presentation of statistical data. Besides, presenting the data in the form of tables, data can also be presented in the form of diagrams. In this chapter, we have discussed One-dimensional diagrammatic presentation of the data. How to draw different types of bar diagrams. A simple bar diagram represents one value whereas the multiple bar diagram represents more than one value. A sub-divided bar diagram represents the different components of a given variable and can also be prepared on percentage basis.

Two-dimensional diagrams are classified as rectangles, squares, circles and pie diagrams. For preparing a rectangle, its length and width are to be determined. For preparing squares and circles, the square root of the given

data is to be calculated and then the side (in case of a square) or the radius (in case of a circle) is determined. A pie diagram is segmented circle, where the segments are determined on the basis of 360" around a point. How to draw Pictograms and Cartograms for the pictorial representations

2.14 REFERENCES

Books:

1. Gupta, S.P. and M.P. Gupta, 2000. Business Statistics, Sultan Chand & Sons: New Delhi.
2. Sinha, S.C. and Dhiman, A.K. 2002. Research Methodology, Vol. 1. Ess Ess Publication, New Delhi.
3. George Argyrons. 2000. Statistics for Social and Health Research with a Guide to SPSS. Sate Publications. New Delhi.

Websites:

https://www.embibe.com/exams/diagrammatic-representations/

https://egyankosh.ac.in/bitstream/123456789/12275/1/Unit-7.pdf

https://www.slideshare.net/VarunPremVaru/diagrammatic-and-graphical-representation-of-data

https://egyankosh.ac.in/bitstream/123456789/20422/1/Unit-14.pdf

https://www.slideshare.net/infinityrulz/module-3-3053101

https://egyankosh.ac.in/bitstream/123456789/13527/1/Unit-8.pdf

2.15 EXERCISE

Exercise 1: The profit for XYZ Company if given below. Represent the data by a simple bar diagram

Year	2010	2011	2012	2013	2014	2015
Profit (%)	30	40	45	56	65	65

Exercise 2: The following data shows expenditure of two families. Represent the data by a subdivided bar diagram.

Item	Family X	Family Y
Food	3500	4000
Clothing	3000	2500
Eduction	3200	3400
Other	2700	2500

Exercise 3: Draw a multiple bar diagram from the following data.

Year	**Sales ('000 Rs)**	**Gross Profi ('000 Rs)**	**Net Profit ('000 Rs)**
2011	100	45	20
2012	115	50	30
2013	120	60	35

Exercise 4: Draw a percentage bar diagram from the following data.

Class	No. of students	Passed students in Mathematics	Passed students in Physics
A	65	34	31
B	60	36	24
C	66	40	26
D	55	26	29

Exercise 5: Represent the following data Rectangle diagram.

Particulars	**2010**	**2011**
Raw Material	320	440
Labour	240	360
Delivery charges	160	280

Exercise 6: The following data relates to the annual plan outlay for particular year for various heads of development. Draw squares for this data.

Heads	**Amount (Rs. in crores)**
Agriculture	2400
Medicines	5800
Bank	13000

Exercise 7: Draw circles for the following data.

Year	No. schools in a city
2000	16
2005	25
2010	81
2015	170

Exercise 8: Draw pie diagram for the following percentage shares of different newspapers sold in Mumbai.

Newspaper	Percentage share
Indian Express	32%
Hindustan Times	25%
Times of Inda	30%
Other	13%
Total	100

Exercise 9: Draw the stem leaf plot for the following data.

74 72 95 89 96 71 87 70 71 90 74 88 65
76 98 72 68 61 86 90 50 54 73 72

❄❄❄❄❄❄❄

3

MEASURES OF CENTRAL TENDANCY

Unit Structure

3.1 OBJECTIVE:

- To get an idea of descriptive statistics in summarization, description and interpretation of data
- To get one single value that describes the characteristics of the entire data.
- To understand the concept of descriptive statistics which has a great significance as it depicts the characteristic of the data as it reduces the entire data into one single value.
- To facilitate comparison, by reducing the whole data into one single value helps in comparison. This can be made either at a point or a period of time.

3.2 REQUISITES OF MEASURE OF CENTRAL TENDENCY:

Since a central tendency is a single value representing a group of values, it is necessary that such a value satisfies the following properties:

1. It should be easy to understand- As statistics is used to simplify the complexity of data average should be easily understood
2. It should be simple to compute- It should be easy to understand and compute so that it can be used widely.

3. It should be based on all the items-The average should be depend on all the items of the data so that even if one observation is dropped then the average changes.

4. It should not be excessively affected by the extreme values- Extreme values may alter the average and reduce its usefulness.

5. It should be rigidly defined and accomplished of further algebraic treatment.

3.3 MEASURE OF CENTRAL TENDENCY:

The various measures of central tendency or averages can be classified into following categories:

1. Mathematical Averages:
   - (a) Arithmetic Mean or Simple Mean
     - Simple
     - Weighted
   - (b) Geometric Mean
   - (c) Harmonic Mean
2. Averages of Position
   - (a) Median
   - (b) Quartiles
   - (c) Deciles
   - (d) Percentiles
   - (e) Mode

Different Notations used:

N= Total number of observations of the population

n= number of observations of the sample

l= lower limit of the class interval

m_i= midpoint or the class mark of the ith class of the data set

h= class width of the class interval

cf= cumulative frequency

$\sum$ = sum of all values of the observations (read as sigma)

3.4 MATHEMATICAL AVERAGES:

Depending on the nature of the data available various mathematical averages of a data set are classified. The data can be of ungrouped type (raw or unclassified) or grouped type (classified).

3.4.1 Arithmetic Mean of Ungrouped type or raw data:

There are two methods to calculate the arithmetic mean for ungrouped data

(i) Direct Method

(ii) Indirect Method or Short-cut method

Direct Method:

In this method we add together the various values of the variable and divide the total by the number of items or observations:

Thus if $x_1 + x_2 + x_3 \ldots x_N$ represents the values of the observations, then the arithmetic mean (A.M) for a population of N observation is

Population mean $\mu = \frac{x_1+x_2+x_3 \ldots x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$ (3-1)

However for a sample containing n observation the A,M is written as

Sample mean $\bar{x} = \frac{x_1+x_2+x_3 \ldots \ldots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (3-2)

The above two formula given in equation (3-1) and (3-2) have different denominators as in statistical analysis the upper case letter N is used to indicate the number of observations of population, while the lower letter n is used for sample observations.

Example 3.1: The following is the monthly income (in thousands) of 12 families in a city:

280, 180, 96, 98, 114, 75, 80, 94, 100, 75, 700, 200. Find the arithmetic mean.

Solution: Applying the formula (3-2) we have

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$= \frac{1}{12}[280 + 180 + 96 + 98 + 114 + 75 + 80 + 94 + 100 + 75 + 700 + 200]$

$= \frac{1}{12}[2092]$

$= 174.33$

Thus, the average income is Rs. 174.33 thousand per month.

Example 3.2: In a survey 5 mobile companies, earned profit (in lakhs) during a year as follows, 25, 20, 10, 35, and 32. Find the arithmetic Mean of the profit earned.

Solution: Applying the formula (3-2) we have

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{1}{5}[25 + 20 + 10 + 35 + 32]$$

$$=24.4$$

Thus, the average profit earned by these mobile companies during a year is Rs. 24.4 lakh.

Alternative formula: In general when observations xi (i=1,2,…n) are grouped as a frequency distribution, then A.M formula (3-2) should be modified as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i \qquad (3\text{-}3)$$

where f_i represents the frequency (number of observations) with which variable x_i occurs in the given data set.

Example 3.3: From the following data of the marks obtained by 60 students of the class calculate the arithmetic mean

Marks	20	30	40	50	60	70
No of Students	5	10	30	5	6	4

Solution:

Let the marks be denoted by X and number of students by f

Marks X	No of students f	fX
20	5	100
30	10	300
40	30	1200
50	5	250
60	6	360
70	4	280

	N=60	$\Sigma fX = 2490$

Using Equation (3-3) $\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{2490}{60} = 41.5$

Hence the average marks of the students is 41.5.

Example: 3.4: If X, Y, Z, W are four chemical substances costing Rs. 25, Rs. 15, Rs. 8 and Rs. 5 per 100 gm respectively and are contained in a given compound in the ratio 1,2, 3 and 4 parts, respectively, then what should be the price of the resultant compound.

Solution: The arithmetic mean is

$$\bar{x} = \sum_{i=1}^{4} f_i x_i = \frac{25\times1+15\times2+8\times3+5\times4}{1+2+3+4} = 9.9$$

Thus, the average price of the resultant compound should be Rs. 9.9 per 100 gm.

INDIRECT METHOD (SHORT-CUT METHOD)

The arithmetic mean can be calculated by using what is known as an arbitrary origin. Suppose we take any figure A as the assumed mean or arbitrary origin and write d as the deviation of the variable X from A as follows:

$$d = x_i - A; \qquad x_i = A + d_i$$

Substituting these values in (3-2) we have

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} xi$$

$$= \frac{1}{n}\sum_{i=1}^{n}(A + d_i) = A + \frac{1}{n}\sum_{i=1}^{n} d_i \qquad (3\text{-}4)$$

If the frequencies of the numeral values are also taken into consideration, then the Equation (3-4) becomes:

$$\bar{x} = A + \frac{1}{n}\sum_{i=1}^{n} f_i d_i \qquad (3\text{-}5)$$

Where n $=\sum_{i=1}^{n} f_i$ total number of observations in the sample.

Example 3.5: The following is the monthly income (in thousands) of 12 families in a town:

280, 180, 96, 98, 104, 75, 80, 94, 100, 75, 600, 200. Find the arithmetic mean.

Solution: The following are the steps to be followed:

Step 1: Take assumed mean

Step 2: Take the deviation of the item from the assumed mean and denote these deviations as d

Step 3: Obtain the sum of these deviation i.e $\sum_{i=1}^{n} d_i$

Step 4: Apply the formula (3-4)

For this question let us take the assumed mean as 150.

Montly Income (in thousands)	d = X-150
280	130
180	30
96	-54
98	-52
104	-46
75	-75
80	-70
94	-56
100	-50
75	-75
600	450
200	50
n = 12	$\Sigma d = 182$

$$\bar{x} = A + \frac{1}{n}\sum_{i=1}^{n} d_i$$

$$= 150+\frac{182}{12}$$

$$= 150+15.17$$

$$=165.17$$

The average income is 165.17 thousand per month

Example 3.6: The daily earning (in rupees) of employees working in a daily basis in a firm are:

Daily earning (Rs)	100	120	140	160	180	200	220
No of Employees	3	6	10	15	20	40	75

Calculate the average daily earning for all employees.

Solution: Let the assumed mean be A=160

Daily Earnings(in Rs) (x_i)	No of Employees (f_i)	$d_i = x_i - A$ $x_i - 160$	$f_i\, d_i$
100	3	-60	-180
120	6	-40	-240
140	10	-20	-200
160	15	0	0
180	20	20	400
200	40	40	1600
220	75	60	4500
Total	169		5880

The required Arithmetic mean is given by

$$\bar{x} = A + \frac{1}{n}\sum_{i=1}^{7} f_i\, di$$

$= 160 + \frac{5880}{169}$

= Rs. 194.79

3.4.2: Arithmetic Mean of Grouped (or Classified) Data

Arithmetic Mean for grouped data can also be calculated by applying any of the following methods:

(i) Direct Method

(ii) Indirect or Short-cut method

The following assumptions are made when calculating arithmetic mean for grouped data:

a) The class interval should be closed (Exclusive)

b) The intervals for class should be equal

c) The values of each observation in each class interval must be uniformly distributed between its lower and upper limits

d) It is assumed that the mid-value of each interval represents the average of all values in that interval, i.e., that all observations are evenly distributed between the lower and upper limits.

Direct Method: The formula is same as equation (3-3) except that x_i is replaced by m_i the mid point of class intervals. Hence, we get

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i m_i, \text{where, } n=\sum f_i \qquad (3\text{-}6)$$

Example 3.7: A company is planning to plant safety measures, the accident data for past 50 weeks are complied and grouped into frequency distribution as shown below. Calculate the arithmetic mean of number of accidents per week.

Number of accidents	0-4	5-9	10-14	15-19	20-24
Number of weeks	2	7	12	17	12

Solution: To find the arithmetic mean we use equation (3-6).

Number of Accidents	Mid-Value (m_i)	Number of Weeks (f_i)	$f_i m_i$
0-4	2	2	4
5-9	7	7	49
10-14	12	12	144
15-19	17	17	289
20-24	22	12	264
TOTAL		n =50	750

The arithmetic mean of number of accidents per week is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i m_i$$

$$= \frac{750}{50}$$

= 15 accidents per week.

Example 3.8: Compute Arithmetic mean by direct method

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No of Students	5	10	25	30	20	10

Solution:

Marks	No of Students f_i)	Mid values (m_i)	$f_i m_i$
0-10	5	5	25
10-20	10	15	150
20-30	25	25	625
30-40	30	35	1050
40-50	20	45	900
50-60	10	55	550
TOTAL	n= 100		3300

The arithmetic mean marks is obtained from

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i m_i$$

$$= \frac{3300}{100}$$

$$=33$$

Short-Cut Method (Step Deviation Method):

When shot cut method is used arithmetic mean is computed by applying the following formula:

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i\, di \times h}{n} \qquad (3\text{-}7)$$

Where A= assumed mean for the data

h = width of the class interval

m_i =mid value of the ith class interval

$di = \frac{mi-A}{h}$ deviation from the assumed mean

Steps: a) Take the assumed mean

b) From the mid point of each class deduct the assumed mean

c) Multiply the respective frequencies of each by these deviations and obtain the total $\sum fidi$

Example 3.9: Calculate the arithmetic mean by the short cut method for the following data, which gives the opinion of 200 people who were interviewed by polling agents.

Age groups (years)	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Number of people	32	42	46	50	20	4	4	2

Solution: Let the assumed mean be A= 45, h=10 (common difference or class width)

Age groups (years)	Fi	mi	$di = \frac{mi - A}{h}$	$fidi$
10-20	32	15	-3	-96
20-30	42	25	-2	-88
30-40	46	35	-1	-46
40-50	50	45	0	0
50-60	20	55	1	20
60-70	4	65	2	8
70-80	4	75	3	12
80-90	2	85	4	8
TOTAL	200			-178

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i\, di \times h}{n}$$

$= 45 + \frac{(-178)\times 10}{200}$

= 36.1 years

Example 3.10: The following distribution gives the pattern of overtime work done by 100 workers of a company. Calculate the average overtime work done per worker.

Overtime hours:	10-15	15-20	20-25	25-30	30-35	35-40
Number of employees:	10	20	36	20	6	8

Solution: Let the assumed mean be A=22.5 and h=5

Overtime (hrs)	Number of employees (fi)	Mid-value (mi)	$di = \frac{mi - A}{h}$	$fidi$
10-15	10	12.5	-2	-20
15-20	20	17.5	-1	-20
20-25	36	22.5	0	0
25-30	20	27.5	1	20
30-35	6	32.5	2	12
35-40	8	37.5	3	24
TOTAL	100			16

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i\, di \times h}{n}$$

$$= 22.5 + \frac{16 \times 5}{100}$$

$$= 23.3 \text{ hours}$$

3.5 ADVANTAGES AND DISADVANTAGES OF ARITHMETIC MEAN:

3.5.1 Advantages:

The arithmetic mean is the most commonly used in practice for the following reasons:

1. The calculation of the arithmetic mean is simple and unique, which means each data set has a unique mean; neither the arraying of data to calculate the median nor the grouping of data to calculate the mode is necessary to calculate the mean.

2. The calculation of Arithmetic mean is based on all observations and is affected by the value of every item in the series.

3. As a single value, the arithmetic mean reflects all the values in the data set.

4. The arithmetic mean is least affected by a fluctuation in sample size. Specifically, its values, which are derived from samples drawn from a population, vary by the smallest amount possible.

5. Being determined by a rigid formula, it lends itself to subsequent algebraic treatment better than median or mode. Some of the algebraic properties of arithmetic mean are as follows:

a) The algebraic sum of deviation of the observations x_i (i=1, 2,…..n) from the arithmetic mean is always zero, that is,

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = \sum_{i=1}^{n} x_i - n\left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i = 0$$

Here the difference $x_i - \bar{x}$ is always referred to as deviation from the arithmetic mean. This result is also true for grouped data.

b) The sum of squares of deviations of all values from arithmetic mean is less than the sum of squares of observations from any other quantity.

This property of Arithmetic mean is also known as the least square property and will be very useful in defining the concepts of standard deviation in the next unit.

c) It is possible to calculate the combined arithmetic mean of two or more data of the same nature, which will be explain in the next chapter.

d) While compiling the data for calculating arithmetic mean, it is possible that we may wrongly read or write certain number of observations, in such cases the correct values can be calculated, which we are going to study in the next chapter.

6. It is calculated value and not based on position in the series.

7. The mean is useful for performing statistical procedures such as comparing the means from several data sets.

3.5.2 Disadvantages:

1. Since the values of mean depends on each and every item of the series, extreme items, i.e, very small and very large items, unduly affect the value of the average. For example, if in a tutorial group there are 4 students and their marks in a test are 65, 75, 10, and 80 the average marks would be

$65+75+10+80 = \frac{230}{4}$

$= 57.5$

One single item, 10 has reduced the average marks considerably. The smaller the number of observations, the greater is likely to be the impact of extreme values.

2. If the class intervals are unequal and open-ended, the arithmetic mean cannot be calculated accurately either at the start or end of the given frequency distribution.

3. The calculation of arithmetic mean sometime become difficult as every data element is used in the calculation (unless the short cut method for grouped data is used to calculate the mean). Moreover, the value so obtained may not be among the observations included in the data.

4. There is no way to calculate the mean for qualitative qualities like intelligence, honesty, beauty, or loyalty.

5. It is not always accurate to use the arithmetic mean for assessing central tendency. The mean provides a characteristic value in that it indicates where most of the values lie, but only when the distribution of the variable is reasonably normal (bell shaped). In case of the U-shaped distribution the mean is not likely to serve a useful purpose.

3.6 CHAPTER END EXERCISES:

1. Explain the term average? What are the merits of a good average?
2. What are the different measures of central tendency? Which is good?
3. Write the advantages and Disadvantages of arithmetic mean?
4. What are the requisites of a measure of central tendency?
5. What are the different types of averages and why arithmetic mean is most commonly used amongst them?
6. Explain with examples the method to calculate the mean using indirect method.

SELF PRACTICE PROBLEMS:

7. During the first 5 months an investor bought the shares at a price of Rs. 100, Rs. 120, Rs. 150, Rs. 200 and Rs. 240 per share. After 5 months what is the average price paid for the shares by him?

8. Salary paid by the company to its employees is as follows:

Designation	Monthly salary (in Rs.)	Number of persons
Senior Manager	35000	1
Manager	30000	20
Executives	25000	70
Jr. Executives	20000	10
Supervisors	15000	150

Calculate the arithmetic mean of salary paid both by direct method and short-cut method.

9. Find the mean by the short cut method:

Weight (in gms) 410-419 420-429 430-439 440-449 450-459 460-449

Number Of Mangoes: 10 20 42 54 45 18

10. Given the following frequency distribution calculate the mean

Class:	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency:	185	77	34	180	136	23	50

3.7 SUGGESTED PRACTICALS:

1. Calculate mean for the following data:

X:	25	26	27	28	29	30	31	32
Frequency:	8	10	9	6	5	4	4	1

2. Compute mean for the following data using indirect method

Weight:	30-40	40-50	50-60	60-70	70-80
Frequency:	6	10	24	40	20

3. Calculate the mean using direct method:

Time:	10-15	15-20	20-25	25-30	30-35	35-40
No of workers	24	42	54	75	45	18

4. Find mean monthly bill using direct method

Electricity bill:	0-5	5-10	10-15	15-20	20-25
Frequency:	2	8	12	23	25

5. Find the mean of first 20 natural numbers.

3.8 REFERENCES:

1. Statistical Methods – S.P Gupta, 8th edition, sultan publisher.

2. Business Statistics- J. K Sharma, Pearson Education.

3. Statistics for Management, Richard I. Levin, 4th edition, Prentice Hall.

❄❄❄❄❄❄❄

4

GEOMETRIC MEAN AND HARMONIC MEAN

Unit Structure

4.0 OBJECTIVE

After completion of the chapter we will able to find,

- Geometric mean
- Harmonic mean
- Relation between Arithmetic, Geometric and Harmonic mean

4.1 SPECIAL TYPES OF PROBLEMS AND SOLUTIONS

Case (i): When frequencies are given in cumulative form (less than cumulative frequency or greater than frequency)

As the more than cumulative frequency is calculated by adding the frequencies from bottom to top, so the first-class interval has the highest cumulative frequency and it goes in decreasing order, but in case of less than frequency the cumulation is done downwards so that the first-class interval has the lowest cumulative frequency and it goes in increasing order.

In either of the cases, we first convert the class interval to inclusive or exclusive type, then the calculation of mean is done in the usual manner.

Example 4.1: Following is the cumulative frequency distribution of the preferred length of kitchen slabs from the preference study on housewives. Find the mean length of slabs.

Length (in meters) more than:	1.0	1.5	2.0	2.5	3.0	3.5
Preference of housewives :	50	48	42	40	10	5

Solution: The given data is converted into exclusive type; the frequency of each class has been found by deducting the given cumulative frequency from the cumulative frequency of the previous class.

CONVERSION INTO EXLUSIVE CLASS

Length	Preference of housewives more than	Class interval	Frequency
1.0	50	1.0-1.5	(50-48) =2
1.5	48	1.5-2.0	(48-42) =6
2.0	42	2.0-2.5	(42-40) =2
2.5	40	2.5-3.0	(40-10) =30
3.0	10	3.0-3.5	(10-5) =5
3.5	5		

Calculation of Mean Length

Class Interval	Mid-value (m_i)	Preference of housewives (f_i)	($m_i x f_i$)
1.0-1.5	1.25	(50-48) =2	2.5
1.5-2.0	1.75	(48-42) =6	10.5
2.0-2.5	2.25	(42-40) =2	4.5
2.5-3.0	2.75	(40-10) =30	82.5
3.0-3.5	3.25	(10-5) =5	16.25
TOTAL		45	116.25

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i m_i$$

$$= \frac{116.25}{45}$$

$$= 2.5833$$

The mean length of the slab is 2.5833 meters.

CASE 2: Frequencies are not given but have to be calculated from the given data:

Sometimes the frequencies will not be given directly and this has to be calculated indirectly. Let's look at an example below.

Example:4.2: 170 clothing factories have the following distribution of average number of workers in various income groups:

Income groups:	750-900	900-1100	1100-1400	1400-1800	1800-2400
Number of firms:	42	32	26	28	42
Average number of Workers:	8	12	8	8	4

Find the mean salary paid to the workers.

Solution: Since the total number of workers (i.e. frequencies) working in the different income groups are not given, therefore these have to be determined as shown below:

Income Group (x_i)	Mid-Values(m_i)	Number of firms	Average number of workers	Frequencies (f_i) =Number of firms x Average number of workers	$m_i f_i$
750-900	825	42	8	336	277200
900-1100	1000	32	12	384	384000
1100-1400	1250	26	8	208	260000
1400-1800	1600	28	8	224	358400
1800-2400	2100	42	4	168	352800
			TOTAL	1320	1632400

The required mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} f_i m_i$$

$$= \frac{1632400}{1320}$$

= Rs. 1236.67

Example 4.3: Find the missing frequencies if the arithmetic mean of the data is 11.09 and N= Σf_i = 60

Class	9.3-9.7	9.8-10.2	10.3-10.7	10.8-11.2	11.3-11.7	11.8-12.2	12.3-12.7	12.8-13.2
Frequency	2	5	F1	F2	14	6	3	1

Solution: The calculation of Arithmetic mean is shown below. Let the assumed mean be A= 11.6

Class	Frequency (f_i)	Mid-values (m_i)	$di = \frac{mi - 11}{0.5}$	$f_i d_i$
9.3-9.7	2	9.5	-3	-6
9.8-10.2	5	10	-2	-10
10.3-10.7	F1	10.5	-1	-F1
10.8-11.2	F2	11.0	0	0
11.3-11.7	14	11.5	1	14
11.8-12.2	6	12	2	12
12.3-12.7	3	12.5	3	9
12.8-13.2	1	13	4	4
TOTAL	60			23-F1

Applying the formula we have:

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i\, di \times h}{n}$$

$$11.09 = 11.0 + \frac{(23 - F1) \times 0.5}{60}$$

$$0.09 = \frac{23 - F1}{120}$$

F1= 23-10.80

F1=12.2

Since the total frequency is 60 and F1=12.2, the other missing frequency

F2 = 60-(2+5+12.2+14+6+3+1)

= 16.8

Case 3: Complete Data are not given:

Example 4.4: The pass result of 50 students who took a class test is given below.

Marks:	40	50	60	70	80	90
Number of Students:	8	10	9	6	4	3

If the mean marks for all the students was 51.6. Find out the mean marks of the students who failed.

Solution: The marks obtained by 40 students who passed are given below

Marks (xi)	Frequency (fi)	$f_i xi$
40	8	320
50	10	500
60	9	540
70	6	420
80	4	320
90	3	270
Total	40	2370

Total marks of all students = 50 x 51.6 = 2580

Total Marks of 40 students who passed = $\Sigma f_i xi = 2370$

Thus marks of the remaining 10 students = 2580-2370 = 210

Hence, the average marks of 10 students who failed are 210/10=21 marks.

Case 4: Incorrect values have been used for calculation of arithmetic mean

Example 4.5: (a) The average dividend declared by a group of 10 chemical companies was 18 percent. Later on, it was discovered that one correct figure, 12 was misread as 22. Find the correct average dividend.

(b) The mean of 200 observations was 50. Later on, it was found that two observations were misread as 92 and 8 instead of 192 and 88. Find the correct mean.

Solution:

(a) Given n=10 and $\bar{x}$=18 percent. We know that

$$\bar{x} = \frac{\sum x}{n} \text{ or } \sum x = n\bar{x} = 10 \text{ x } 18 = 180$$

Since one number was misread as 22 instead of 12 we therefore subtract the incorrect value and add the correct value to above equation, Thus we have the correct total as

$\sum x$ = 180-22+12 = 170

Hence the correct mean is

$$\bar{x} = \frac{\sum x}{n} = 170/10$$

=17 percent

(b) Given that n=200, $\bar{x}$=50, we know that

$$\bar{x} = \frac{\sum x}{n} \text{ or } \sum x = n\bar{x} = 200x50 = 10000$$

Since two observations were misread, therefore the correct total can be obtained by

$$\sum x = 10000 - (92 + 8) + (192 + 88) = 10180$$

Hence the correct mean is:

$$\bar{x} = \frac{\sum x}{n} = \frac{10180}{200} = 50.9$$

Combined Mean:

It is possible to calculate the combined or pooled arithmetic mean of twoor more than two sets of data of the same nature.

Let $\bar{x}1\ and\ \bar{x}2$ be arithmetic mean of two sets of data of the same nature of size n1 and n2 respectively. Then their combined arithmetic mean can be calculated as:

$$\bar{x}_{12} = \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2} \quad (4.1)$$

This result can be generalized in the same way for more than two sets of data of different sizes having different arithmetic means.

Example 4.6: There are two units of an automobile company in two different cities employing 760 and 800 persons, respectively. The arithmetic means of monthly salaries paid to persons in these two units are Rs. 18750 and Rs. 16950 respectively. Find the combined arithmetic mean of salaries of the employees in both the units.

Solution: Let $n_1 = 760$ and $n_2 = 800$ be the number of persons working in unit 1 and 2 respectively and $\bar{x}_1 = 18750$ and $\bar{x}_2 = 16950$

Thus, the combined mean of salaries paid by the company is:

$$\bar{x}_{12} = \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2} = \frac{760x18750+800x16950}{760+800} = Rs.17826.92 \text{ per month}$$

Example 4.7: The arithmetic mean weight of 50 men is 55kgs and mean of 25 women is 45kgs. Find the mean weight of the combined group.

Solution: Let $n_1 = 50$ and $n_2 = 25$ be the number of men and women respectively and $\bar{x}_1 = 55$ and $\bar{x}_2 = 45$

Thus, the combined mean of the group is

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{50x\ 55\ + 25x45}{50 + 25} = 51.67kg$$

4.2 WEIGHTED ARITHMETIC MEAN:

As the arithmetic mean gives importance to each data observations in the data set. However, there can be a situation where each observation does not have equal importance, in such cases, computing arithmetic mean by using equation (3.1) may not represent the characteristic of data set and thus misleading the learners. In such situation, we attach to each observations a weight say $w_1, w_2, \ldots \ldots w_n$ as an indicator of their importance and compute the weighted mean or average denoted by $\bar{x}_w\ as\ follows$:

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i} \qquad (4.2)$$

When can we use Weighted Arithmetic Mean:

1. When the importance of all observations are not equal.

2. When frequencies of various classes are varying widely.

3. When we average the ratios, percentages or rates.

Example 4.7: An examination was held to decide an award to the students. The weights of various subjects were different. The marks obtained by three students out of 100 are as follows:

Subject	Weight	Student A	Student B	Student C
Physics	3	62	61	67
Chemistry	2	55	53	60
Mathematics	4	60	57	62
English	1	67	77	49

Calculate the weighted arithmetic mean of this award.

Solution: The weighted arithmetic mean is calculated using equation (4.2)

Subject	Weight (w_i)	Student A (x_i)	$x_i\ w_i$ A	Student B (x_i)	$x_i\ w_i$ B	Student C (x_i)	$x_i\ w_i$ C
Physics	3	62	186	61	183	67	201

Chemistry	2	55	110	53	106	60	120
Mathematics	4	60	240	57	228	62	248
English	1	67	67	77	77	49	49
Total	10		603		594		618

Applying equation (4.2) we have

$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i} = \frac{603+594+618}{10}$ =181.5

Example:4.8 Salary paid by a company to its employees is as follows: Calculate the weighted mean.

Designation: Senior Manager Manager Executives Jr. Executives supervisor

Monthly salary: 35000 30000 25000 20000 15000

No of persons: 1 20 70 10 150

Weights: 5 4 3 2 1

Solution: The weighted Arithmetic mean is calculated using equation (4.2)

Designation	Monthly salary (x_i)	Weights ((w_i)	$x_i\, w_i$
Senior Manager	35000	5	175000
Manager	30000	4	120000
Executives	25000	3	75000
Jr. Executives	20000	2	40000
Supervisor	15000	1	15000
TOTAL		15	425000

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i} = \frac{425000}{15} = 28333.33$$

4.3: GEOMETRIC MEAN:

When we deal with quantities that change over a period of time simple average may not work, instead we use an average rate of change to represent the average growth or decline rate in data over a period of time. This gave raise to another measure of central tendency called geometric mean.

Consider for example the geometric mean can be used to find the average percent increase in sales, production, population or other economic or business dealings. Like if the price increase from 1999 to 2001 are 5%, 10% and 18% respectively. The average annual increase is not 11% as given by arithmetic mean but 10.9% as obtained by geometric mean.

Geometric mean is defined as the nth root of the product of N times or values. If there are two items we take square root; if there are three then cube root and so on… symbolically,

G.M= $\sqrt[n]{x_1, x_2, x_3 \ldots x_n}$ where $x_1,\ x_2 \ldots . x_n$ are various items of the series.

When the number of observations are three or more the calculation of root becomes difficult therefore to simplify the calculation logarithms are used. Hence Geometric Mean is calculated as follows:

For discrete Series: $log\ G.M = \frac{\sum log\ x}{N}$ or $G.M = Antilog\ \frac{\sum logx}{N}$ (4.3)

$$G.M = Antilog\ \frac{\sum f\ logx}{N}, \text{N}=\sum f \quad (4.4)$$

For Continuous series: $log\ G.M = \frac{\sum log\ m}{N}$ or

$$G.M = Antilog\ \frac{\sum logm}{N}, \quad where\ m\ is\ the\ midpoint\ of\ each\ class \quad (4.5)$$

$$G.M = Antilog\ \frac{\sum f\ logm}{N}, \quad \text{N}=\sum f \quad (4.6)$$

Example 4.9: Calculate the geometric mean of the following:

X: 257, 475, 5, 8, 9

X	Log X
257	2.4099
475	2.6767
5	0.6990
8	0.9031
9	0.9542
TOTAL	7.6429

Using equation (4.3) we have

$$G.M = Antilog\ \frac{\sum logx}{N}$$

= Antilog (7.6429/5)

= Antilog (1.52858)

= 33.77

Example 4.10: A given instrument is assumed to depreciate 30 percent in first year, 20 percent in second year and 10 percent for the next three years, each percentage being calculated on the diminishing value. What is the average depreciation recorded on the diminishing value for the period of five years?

Solution: We use equation (4.4) to find the geometric mean.

Rate of depreciation (xi)	Number of years (fi)	log (xi)	f log(xi)
30	1	1.4771	1.4771
20	1	1.3010	1.3010
10	3	1	3
TOTAL	5	TOTAL	5.7781

$G.M = Antilog\ \frac{\sum f\ logx}{N}$

= Anitlog (5.7781/5)

= Antilog (1.15562)

=14.309

Hence the average rate of depreciation for first five years is 14.309 percent.

Example 4.11: From the following data calculate the geometric mean.

Marks:	0-10	10-20	20-30	30-40	40-50
No. of students:	8	10	22	6	4

Solution: We use equation (4.6) to compute the geometric mean. we first find the midpoint of the class interval.

Marks	No of students	Midpoints (m)	log m	f x log m
0-10	8	5	0.6990	5.5920
10-20	10	15	1.1761	11.761
20-30	22	25	1.3978	30.7516
30-40	6	35	1.5441	9.2646
40-50	4	45	1.6532	6.6128
TOTAL	50		TOTAL	63.982

$$G.M = Antilog\ \frac{\sum f\ logm}{N}$$

= Antilog (63.982/50)

= Antilog(1.27964)

= 19.0388

4.3.1: Combined Geometric Mean:

The combined geometric mean is pooling the geometric mean of different sets or group of data and is defined as follows:

$$logG.M = \sum_{i=1}^{n} \frac{nilogGi}{n} \qquad (4.7)$$

where G_i is the geometric mean of the i^{th} data set having n_i number of observations.

Example 4.12: Three sets of data contain 5, 6, 9 observations and their geometric mean are 7.75, 6.52 and 12.12 respectively. Find the combined geometric mean of 20 observations.

Solution: Applying Equation (4.7) we can get the combined geometric mean as follows:

$$G.M = Antilog\left[\frac{n_1 log\ G_1 + n_2 log\ G_2 + n_3 log\ G_3}{n_1 + n_2 + n_3}\right]$$

$$= Antilog\left[\frac{5\ log\ (7.75) + 6\ log\ (6.52) + 9\ log\ (12.12)}{20}\right]$$

$$= Antilog\left[\frac{5\ x\ 0.8893 + 6x\ 0.8142 + 9x\ 1.0835}{20}\right]$$

= Antilog [19.0832/20]

= 8.9982

Hence the combined mean of 20 observations is 8.9982.

Example 4.13: Find the combined geometric mean of three observations that contain 5, 8, 10 observations and their respective geometric mean are 6.7205, 8.6905, 9.1456.

Solution: Applying Equation (4.7) we can get the combined geometric mean as follows:

$$G.M = Antilog\left[\frac{n_1 log\ G_1 + n_2 log\ G_2 + n_3 log\ G_3}{n_1 + n_2 + n_3}\right]$$

$$= \text{Antilog}\left[\frac{5\ log\ (6.7205) + 8\ log\ (8.6905) + 10\ log\ (9.1456)}{23}\right]$$

$$= \text{Antilog}\left[\frac{5\ x\ 0.8274 + 8x\ 0.9390 + 10x\ 0.9612}{23}\right]$$

= Antilog [21.261/23]

= Antilog(0.92439)

= 8.4021

Hence the combined mean of 23 observations is 8.4021

4.3.2: Weighted Geometric Mean

If each observations $x_i (i = 1,2,\ldots n)$ are given different weights or importance say w_i (i=1,2,…n) respectively, then their weighted geometric mean is defined as:

$$\text{G.M (w)} = \text{Antilog}\left[\left(\frac{1}{n}\right)\sum wlogx\right]$$

$$= \text{Antilog}\left[\left(\frac{1}{\sum\omega}\right)\sum wlogx\right] \quad (4.8)$$

Examples 4.14: The weighted geometric mean of the four numbers are 20, 18, 12, 14 is 11.75. If the weights of the first three numbers are 1, 3 and 4 respectively, find the weights of the fourth number.

Solution: Let the weight of the fourth number be w. Then the weighted geometric mean of the four numbers can be calculated as shown below using equation (4.8)

x	Weight of each (w)	logx	wlogx
20	1	1.3010	1.3010
18	3	1.2553	3.7659
12	4	1.0792	4.3168
4	W	0.6021	0.6021w
TOTAL	8+w		9.3837+0.6021w

This the weighted G.M is

$$\text{G.M.} = \text{Antilog}\left[\left(\frac{1}{\sum\omega}\right)\sum wlogx\right]$$

$$\log(11.75) = \left[\left(\frac{1}{8+w}\right)(9.3837 + 0.6021w)\right]$$

(8+w) (1.0700) =(9.3837+0.6021w)

8.56+1.0700w = 9.3837+0.6021w

$$0.4679w = 0.8237$$

$$w = 1.7604$$

Example 4.15: The weighted geometric mean for the four numbers are 8, 25, 17, 30 is 12.3. If the weights of the first three numbers are 5, 3, and 4 respectively, find the weight of fourth number.

Solution: Let the weight of the fourth number be w. Using equation (4.8) we have

x	Weight of each (w)	logx	wlogx
8	5	0.9031	4.5155
25	3	1.3979	4.1937
17	4	1.2304	4.9216
3	W	0.4771	0.4771w
TOTAL	12+w		13.6308+0.4771w

This the weighted G.M is

$$\text{G.M.} = \text{Antilog}\left[\left(\frac{1}{\sum\omega}\right)\sum wlogx\right]$$

$$\log(12.3) = \left[\left(\frac{1}{12+w}\right)(13.6308 + 0.4771w)\right]$$

$$(12+w)\,(1.0899) = (13.6308+0.4771w)$$

$$13.0788+1.0899w = 13.6308+0.4771w$$

$$0.6128w = 0.552$$

$$w = 0.901$$

4.3.3: Advantages, Disadvantages and Applications of G.M.:

Advantages:

(a) The value of geometric mean is computed by taking into account all the values of the observation, hence it is rigidly defined.

(b) It can be used to calculate the average ratio and percentage, as well as the rate of increase and decrease.

(c) The value of geometric mean is smaller or equal to arithmetic mean the reason is, it gives more weight to large items and less weight to small ones then does the arithmetic mean. Sometimes it may turn out to be same as arithmetic mean, usually it is smaller.

(d) It is capable of algebraic manipulations.

Disadvantages:

(a) It is difficult to compute and interpret.

(b) When we have negative or zero observations geometric mean cannot be calculated.

(c) Geometric mean has very restricted applications because of the above reasons.

Applications:

(a) The geometric mean is used in the construction of index numbers.

(b) In application of social economic study where smaller observations are given more importance geometric mean is useful as its value is smaller than arithmetic mean.

4.4: Harmonic Mean:

The harmonic mean (H.M) of set of observations is defined as the reciprocal of Arithmetic Mean of the reciprocal of the individual observations, that is,

$$\frac{1}{H.M} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$$

Or
$$H.M = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)} \text{ (for ungrouped data)} \qquad (4.9)$$

For a grouped data, when $f_1, f_2, \ldots f_n$ are the frequencies attached to $x_1, x_2, \ldots . x_n$ the H.M is given by

$$H.M = \frac{n}{\sum_{i=1}^{n} f_i\left(\frac{1}{x_i}\right)} \qquad (4.10)$$

Where n =$\sum_{i=1}^{n} f_i$

Example 4.16: Calculate harmonic mean of the following:

1, 0.5, 10, 35, 125, 0.1, 4.0, 11.2

Solution: We use equation (4.9) to calculate the harmonic mean

X	1/X
1	1
0.5	2
10	0.1
35	0.0286

125	0.008
0.1	10
4.0	0.25
11.2	0.0893
TOTAL	13.4759

$$H.M = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

H.M= $\frac{8}{13.4759}$

= 0.5937

Example 4.17: From the following data compute the value of harmonic mean:

Marks: 10 15 30 40 50

No of Students: 20 30 50 15 5

Solution: Using Equation (4.10) we will calculate the Harmonic Mean.

Marks (X)	f	Reciprocal (1/X)	f. (1/X)
10	20	0.1	2
15	30	0.0666	1.998
30	50	0.0333	1.665
40	15	0.025	0.375
50	5	0.02	0.1
TOTAL	120		6.138

$$H.M = \frac{n}{\sum_{i=1}^{n} f_i\left(\frac{1}{x_i}\right)}$$

$= \frac{120}{6.138}$

= 19.550

4.4.1: Advantages, Disadvantages of Harmonic Mean:

Advantages

(a) Every element is involved in the computation of H.M

(b) As the reciprocal is taken in the calculation of Harmonic Mean, more importance is given to smaller values in the data set.

(c) Harmonic Mean can be used for further algebraic analysis.

Disadvantages:

(a) If the observations have negative and zero elements then H.M cannot be calculated.

(b) It cannot be used for business problems.

(c) Since the calculation is complicated, it does not represent the characteristic of the data set.

4.4.2: Applications of Harmonic Mean:

The Harmonic mean has lot of restrictions. It is helpful in computing the typical rate of increase of profits of a priority or average speed at that a journey has been performed, or the typical value at that the article has been sold out. The rate usually indicates the relation between two different types of measuring units that can be expressed reciprocally.

4.4.3 Relationship between A.M, G.M and H.M

For any set of observations, it's A.M, G.M and H.M has the following order.

$$A.M \geq G.M \geq H.M$$

The sign of "=" holds if and only if all the observations are identical.

Therefore, Arithmetic Mean is greater than Geometric Mean and Geometric Mean is greater than Harmonic Mean.

4.5 SOME SUGGESTED PRACTICAL:

Conceptual Questions:

1. Distinguish between simple and weighted average and state the circumstances under which the latter should be employed.
2. Define simple and weighted geometric mean and state under what circumstances they are recommended to use.
3. Discuss the advantages, disadvantages of geometric mean.
4. Discuss the advantages and disadvantages of harmonic mean.
5. Write the application of both geometric and harmonic mean.

4.6 SELF PRACTICE PROBLEMS:

1. The mean monthly salary paid to all employees in a company is Rs. 16000. The mean monthly salaries paid to technical and non-technical employees are Rs. 18000, and Rs. 12000 respectively. Determine the percentage of technical and non-technical employees in the company.
2. The mean marks in statistics of 100 students in a class was 72 percent. The mean marks of boys was 75 percent, while their number was 70 percent. Find out the mean marks of girls in the class.

3. The mean of 200 items was 50. Later on it was discovered that two items were misread as 92 and 8 instead of 192 and 88. Find out the correct mean.

4. Find the Harmonic mean of the following observations:

Class interval:	0-10	10-20	20-30	30-40
Frequency:	5	8	3	4

5. Calculate the A.M, G.M and H.M of the following observations and show that A.M>G,M>H.M.

32 35 36 37 39 41 43

6. Calculate the Harmonic mean of the profit earned.

Profit:	20	21	22	23	24	25
No of companies:	4	2	7	1	3	1

7. Find the geometric mean of the following:

Dividend:	0-10	10-20	20-30	30-40	40-50
No of Companies:	5	7	15	25	8

8. The weighted geometric mean of four numbers 21, 19, 13 and 14 are 10.52. If the weight of the first three are 1, 4, 5 respectively. Find the weight of the fourth number.

9. The arithmetic mean height of 50 students of a college is 5'8". The height of 30 of them are 5'6". Find the arithmetic mean height of the remaining 20 students.

10. The mean monthly salary paid to 100 employees of a company was Rs. 5000. The mean monthly salaries paid to male and female employees were Rs. 5200 and Rs. 4200 respectively, Determine the percentage of males and females employed by the company.

11. Find the missing frequency if the mean of the observation is 27.25.

Classes:	0-10	10-20	20-30	30-40	40-50
Frequency:	5	20	22	18	-

12. A batsman plays four one day matches. His bating score in each of these matches is 40, 60, 30, and 50. If the weights assigned to these matches are 4,3,2,1 respectively find his average score.

4.7 REFERENCES:

1. Statistical Methods – S.P Gupta, 8th edition, sultan publisher.

2. Business Statistics - J. K Sharma, Pearson Education.

3. Statistics for Management, Richard I. Levin, 4th edition, Prentice Hall.

❄❄❄❄❄❄❄

5

AVERAGES

Unit Structure

5.0 INTRODUCTION

What is the positional average?

Positional averages are those averages whose values are worked out on the basis of their position in the statistical series.

Averages of Position:

a) Median

b) Quartiles

c) Deciles

d) Percentiles

e) Mode

Various methods of calculating mathematical averages of a data set are classified in accordance of the nature of data available that is, ungrouped (unclassified or raw) or grouped (classified) data.

5.1 OBJECTIVE

- To get an idea of descriptive statistics in summarization, description and interpretation of data
- To get one single value that describes the characteristics of the entire data.

- To compute arithmetic mean, geometric mean, harmonic mean, median, mode, quartiles, percentiles, and deciles.
- To understand the concept of descriptive statistics which has a great significance as it depicts the characteristic of the data as it reduces the entire data into one single value.
- To facilitate comparison, by reducing the whole data into one single value helps in comparison. This can be made either at a point or a period of time.

5.2 MEDIAN, ADVANTAGES, DISADVANTAGES AND APPLICATIONS OF MEDIAN.

Median: The Median is also a measure of central tendency. Unlike the arithmetic mean, this median is based on the position of a given observation in a series arranged in an ascending or descending order. Therefore, it is called a positional average. Sequence of data in the sense that half of the observations are smaller and half are larger than this value. The median is thus a measure of the location or centrality of the observations.

(i) Ungrouped Data :

The Median is the middle observation of an ordered (Ascending or Descending) data set.

1. When n is odd: The size of $\left(\frac{n+1}{2}\right)$ th observations.
2. When n is even: there are two middle values. They are arithmetic mean of $\left(\frac{n}{2}\right)$th observation and $\left(\frac{n}{2}+1\right)$ th observations.

Example 1:

Find median of the following observations: 1, 2, 4, 7, 9, 5, 3,

Solution: Arrange the observations in an ascending order i.e. 1, 2, 3, 4, 5, 7, 9.

The size of $\left(\frac{n+1}{2}\right)$ th observation, when n is odd. Since n=7, i.e. odd,

The median is the size of $\left(\frac{7+1}{2}\right)$ th, i.e. 4^{th} observation. Hence median is 4.

Example 2:

Find median of the following observations: 8, 9, 1, 6, 5, 7, 2, 3

Solution: Arrange the observations in an ascending order i.e. 1, 2, 3, 5, 6, 7, 8, 9

Here n = 8, i.e even.

Since there are two middle terms and their average is of $\left(\frac{8}{2}\right)$th observation and $\left(\frac{8}{2}+1\right)$ th observations.

i.e. 4th and 5th observations.

Median = $\frac{5+6}{2}$ = 5.5

(ii) Grouped Data:

(a) Discrete Variate (b) Continuous Variate

(a) **Discrete Variate:** Discrete means which is not continuous then the data is already in the order. Here cumulative frequency is computed and the median is determined in a manner similar to that of individual observations.

Example 3: Locate median of the following frequency distribution:

Variable (x):	10	11	12	13	14	15	16
Frequency (f):	8	15	25	20	12	10	5

Solution:

Variable (x):	10	11	12	13	14	15	16
Frequency (f):	8	15	25	20	12	10	5
Cumulative Frequency:	8	23	48	68	80	90	95

Here N = 95, which is odd. Thus, median is size of $\left[\frac{95+1}{2}\right]$ th i.e. 48th observation. From the table 48th observation is 12. Median = 12.

Example 4: Locate median of the following frequency distribution:

Variable (x):	0	1	2	3	4	5	6	7
Frequency (f):	7	14	18	36	51	54	52	20

Solution:

Variable (x):	0	1	2	3	4	5	6	7
Frequency (f):	7	14	18	36	51	54	52	20
Cumulative Frequency:	7	21	39	75	126	180	232	252

Here N = 252, i.e. even. then $\left(\frac{n}{2}\right)$th observation and $\left(\frac{n}{2}+1\right)$ th observations is $\left[\frac{252}{2}\right]$ = 126 and $\left[\frac{252}{2}+1\right]$ = 127.

Median is the mean of the size of 126th and 127th observation. From the table we note that 126th observation is 4 and 127th observation is 5.

Median = $\frac{4+5}{2}$ = 4.5

(b) **Continuous Variate:** While computing the value of median in a continuous variate the middle item is the ($\frac{N}{2}$)th item. Using the cumulative frequencies we can locate the class interval in which median lies. But the exact value has to be calculated by following formula:

Median M = $l_1 + \frac{(l_2 - l_1)\left[\frac{N}{2} - c.f\right]}{f}$

Let l_1 = Lower limit of median class

l_2 = Upper limit of median class

f = frequency of the median class

c.f = cumulative frequency of the pre median class.

Example 5: Calculate median for the following data:

Height in inches: 3-4 4-5 5-6 6-7 7-8 8-9 9-10 10-11

No. of saplings: 3 7 12 16 22 20 13 7

Solution:

Class Intervals	Frequency (f)	Less than type cumulative frequency (c.f)
3-4	3	3
4-5	7	10
5-6	12	22
6-7	16	38
7-8	22	60
8-9	20	80
9-10	13	93
10-11	7	100
Total	Σ f=100	

N= Σ f = 100

$\frac{N}{2} = \frac{100}{2}$ = 50, the median class is 7-8. l_1 = 7 , l_2 = 8, f = 22, c.f = 38

Median M = $l_1 + \frac{(l_2 - l_1)\left[\frac{N}{2} - c.f\right]}{f}$

Substitute the values in the given formula M =7 + $\frac{(8-7)[50-38]}{22}$ = 7.55 inches

Example 5: In a factory employing 3000 persons, 5 per cent earn less than Rs. 150 per day, 580 earn from Rs. 151 to Rs. 200 per day, 30 per cent earn from Rs. 201 to Rs. 250 per day, 500 earn from Rs. 251 to Rs. 300 per day, 20 per cent earn from Rs. 301 to Rs. 350 per day, and the rest earn Rs. 351 or more per day. What is the median wage?

Solution: Calculations of median wage per day are shown in Table

Earnings in (Rs.)	**Percentage of Workers (Per cent)**	**Number of Persons (f)**	**Cumulative Frequency (c.f)**
Less than 150	5	150	150
151–200	—	580	730
201–250	30	900	1630 ← Median class
251–300	—	500	2130
301–350	20	600	2730
351 and above	—	270	3000
		N= 3000	

Median observation = $(\frac{n}{2})$ th = $\frac{3000}{2}$ = 1500th observation.

This observation lies in the class interval 201–250.

Median M = $l_1 + \frac{(l_2 - l_1)\left[\frac{N}{2} - c.f\right]}{f}$

= 201+$\frac{1500-730}{900}$ x 50

= 201 + 42.77 = Rs. 243.77.

Hence, the median wage is Rs. 243.77 per day.

Advantages, Disadvantages, and Applications of Median

Advantages

- It is easy to understand and easy to calculate, especially in series of individual observations and ungrouped frequency distributions.
- Median is unique, i.e. like mean, there is only one median for a set of data.
- Median can also be located graphically.
- The value of median is easy to understand and may be calculated from any type of data. The median in many situations can be located simply by inspection.
- Values in the data set do not affect the calculation of the median value and therefore it is the useful measure of central tendency when such values do occur.
- The median value may be calculated for an open-ended distribution of data set.

Disadvantages

- The median is not capable of algebraic treatment. For example, the median of two or more sets of data cannot be determined.
- The value of median is affected more by sampling variations, that is, it is affected by the number of observations rather than the values of the observations. Any observation selected at random is just as likely to exceed the median as it is to be exceeded by it.
- Since median is an average of position, therefore arranging the data in ascending or descending order of magnitude is time consuming in case of a large number of observations.
- The calculation of median in case of grouped data is based on the assumption that values of observations are evenly spaced over the entire class interval.

Applications

The median is helpful in understanding the characteristic of a data set when

- Observations are qualitative in nature
- Extreme values are present in the data set
- A quick estimate of an average is desired.

E x e r c i s e :

- Define median and discuss its advantages and disadvantages.
- When is the use of median considered more appropriate than mean?

5.3 PARTITION VALUES, QUARTILES, DECILES AND PERCENTILES.

Quartiles divide the data into four equal parts; deciles divide the data into ten equal parts and percentiles divide the data into hundred equal parts. These partition values are used to fragment a distribution into smaller parts which are easier to measure, analyze and understand. The measures of central tendency which are used for dividing the data into several equal parts are called partition values.

All these values can be determined in the same way as median. The only difference is in their location.

Quartiles:

Whenever we have an observation and we wish to divide it, there is a chance to do it in different ways. So, we use the ***median*** when a given observation is divided into two parts that are equal. Likewise, **quartiles** are values that divide a complete given set of observations into four equal parts.

Basically, there are three types of quartiles, first quartile, second quartile, and third quartile. The other name for the first quartile is lower quartile. The representation of the first quartile is 'Q_1.' The other name for the second quartile is median. The representation of the second quartile is by 'Q_2.' The other name for the third quartile is the upper quartile. The representation of the third quartile is by 'Q_3.'

First Quartile is generally the one-fourth of any sort of observation. However, the point to note here is, this one-fourth value is always less than or equal to 'Q_1.' Similarly, it goes for the values of 'Q_2'and 'Q3.'

$$Q_1 = l_1 + \frac{(l_{2-}\, l_1)\left[\frac{N}{4} - c.f\right]}{f}$$

(ii) The second quartile Q_2 has the same number of observations above and below it. It is therefore same as median value.

$$Q_2 = l_1 + \frac{(l_{2-}\, l_1)\left[\frac{N}{2} - c.f\right]}{f}$$

(iii) The quartile Q_3divides the data set in such a way that 75 per cent of the observations have a value less than Q_3 and 25 per cent have a value more than Q_3, i.e. Q_3 is the median of the order values that are above the median.

$$Q_3 = l_1 + \frac{(l_{2-}\, l_1)\left[\frac{3N}{4} - c.f\right]}{f}$$

Deciles:

Deciles are those values that divide any set of a given observation into a total of ten equal parts. Therefore, there are a total of nine deciles. These representation of these deciles are as follows – D_1, D_2, D_3, D_4, …… D_9.

D_1 is the typical peak value for which one-tenth (1/10) of any given observation is either less or equal to D_1. However, the remaining nine-tenths(9/10) of the same observation is either greater than or equal to the value of D_1.

D_1 is calculated using the formula

$$D_1 = l_1 + \frac{(l_2 - l_1)\left[\frac{N}{10} - c.f\right]}{f}$$

l_1 = Lower limit of first decile class

l_2 = Upper limit of first decile class

f = frequency of the first decile class

c.f = cumulative frequency of the class preceding the first decile class.

For kth decile D_K; where k = 1, 2, 3,….9 is located as the one for which cumulative frequency exceeds $\frac{KN}{10}$. Where $D_k = l_1 + \frac{(l_2 - l_1)\left[\frac{KN}{10} - c.f\right]}{f}$

If the third decile D_3 is to be calculated put k=3 in the above formula.

Percentiles: The values of observations in a data when arranged in an ordered sequence can be

divided into hundred equal parts using ninety nine percentiles, Pi (i = 1, 2, . . ., 99).

The number of observations less than P_1 is $\frac{N}{100}$, the number of observations less than P_2 is $\frac{2N}{100}$ …….. The number of observations less than P_k is $\frac{KN}{100}$

Example 6: Find the three quartiles, 3rd and 7th Deciles, 9th and 87th percentiles from the following data:

Daily wages in Rs.	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55
No. of Workers	12	28	36	50	25	18	16	10	5

Solution:

Daily wages in Rs.	No. of Workers (f)	Less than cumulative Frequency (c.f)
10-15	12	12
15-20	28	40
l_1 20-25 l_2	36 f	76 Q_1

l_1 25-30 l_2	50 f	126 Q_2
l_1 30-35l_2	25 f	151 Q_3
35-40	18	169
40-45	16	185
45-50	10	195
50-55	5	200
Total	N=200	

$Q_1 = l_1 + \frac{(l_2 - l_1)\left[\frac{N}{4} - c.f\right]}{f}$ $\frac{N}{4} = \frac{200}{4} = 50,$

$= 20 + \frac{(25-20)(50-40)}{36}$

$= 20 + \frac{(5)(10)}{36}$

$Q_1 = 20+1.3889 = \text{Rs. } 21.39$

$Q_2 = l_1 + \frac{(l_2 - l_1)\left[\frac{N}{2} - c.f\right]}{f}$ $\frac{N}{2} = \frac{200}{2} = 100,$

$= 25 + \frac{(30-25)(100-76)}{50}$

$= 25 + \frac{(5)(24)}{50}$

$Q_2 = 25+2.4 = \text{Rs.}27.4$

$Q_3 = l_1 + \frac{(l_2 - l_1)\left[\frac{3N}{4} - c.f\right]}{f}$ $\frac{3N}{4} = \frac{3\times200}{4} = 150,$

$= 30 + \frac{(35-30)(100-126)}{25}$

$= 30 + \frac{(5)(24)}{25}$

$Q_3 = 30+4.8 = \text{Rs.}34.8$

3rd and 6th Deciles: Calculation of deciles:

Daily wages in Rs.	No. of Workers (f)	Lessthan cumulative Frequency (c.f)
10-15	12	12
15-20	28	40
l_1 20-25 l_2	36 f	76 D_3
25-30	50	126
l_1 30-35l_2	25 f	151 D_7

35-40	18	169
40-45	16	185
45-50	10	195
50-55	5	200
Total	N=200	

$$for\ D_3 = \frac{3N}{10} = \frac{3\times200}{10} = 60,\ D_3 = l_1 + \frac{(l_{2-}\ l_1)\left[\frac{3N}{10} - c.f\right]}{f}$$

$$= 20 + \frac{(25-20)(60-40)}{36}$$

$$= 20 + \frac{(5)(20)}{36}$$

$$= 20 + 2.7778$$

D_3 = Rs.22.78

$$for\ D_7 = \frac{7N}{10} = \frac{7\times200}{10} = 140,\ D_7 = l_1 + \frac{(l_{2-}\ l_1)\left[\frac{7N}{10} - c.f\right]}{f}$$

$$= 30 + \frac{(35-30)(140-126)}{25}$$

$$= 30 + \frac{(5)(14)}{25}$$

$$= 30 + 2.80$$

D_7 = Rs.32.80

9th and 87th percentiles from the following data:

Daily wages in Rs.	No. of Workers (f)	Less than cumulative Frequency (c.f)
10-15	12	12
l_1 15-20 l_2	28 f	40 P_9
20-25	36	76
25-30	50	126
30-35	25	151
35-40	18	169
$l_1$40-45 l_2	16 f	185 P_{87}
45-50	10	195
50-55	5	200
Total	N=200	

For $\frac{9N}{100} = \frac{9\times200}{100} = 18$,

$$P_9 = l_1 + \frac{(l_{2-}\ l_1)\left[\frac{9N}{100} - c.f\right]}{f}$$

$$= 15 + \frac{(20-15)(18-12)}{28}$$

$$= 15 + \frac{(5)(6)}{28}$$

$$= 15 + 1.0714$$

$$P_9 = Rs.16.07$$

For $\frac{87N}{100} = \frac{87\times200}{100} = 174$

$$P_{87} = l_1 + \frac{(l_{2-}\ l_1)\left[\frac{87N}{100} - c.f\right]}{f}$$

$$= 40 + \frac{(45-40)(174-169)}{16}$$

$$= 40 + \frac{(5)(5)}{16}$$

$$= 40 + 1.5625$$

$$P_{87} = Rs.41.56$$

Exercise:

1. What are quartiles of a distribution? Explain their uses.

2. Describe the similarities and differences among median, quartiles, and percentiles as descriptive measures of position.

5.4 GRAPHICAL METHOD FOR CALCULATING PARTITION VALUES.

First we prepare the cumulative frequency table, and then the cumulative frequencies are plotted against the upper or lower limits of the corresponding class intervals. By joining the points the curve so obtained is called a cumulative frequency curve or **Ogive**. There are two types of ogives:

1. **Less than ogive:** Plot the points with the upper limits of the class as abscissae and the corresponding less than cumulative frequencies as

ordinates. The points are joined by free hand smooth curve to give less than cumulative frequency curve or the less than Ogive. It is a rising curve.

2. **Greater than ogive:** Plot the points with the lower limits of the classes as abscissa and the corresponding Greater than cumulative frequencies as ordinates. Join the points by a free hand smooth curve to get the "More than Ogive". It is a falling curve.

Example 7: Draw the two ogives for the following frequency distribution of the weekly wages of (less than and more than) number of workers

Weekly wages	**Number of workers**	**C.F (less than)**	**C.F (More than)**
0 – 20	41	41	201
20 – 40	51	92	160
40 – 60	64	156	109
60 – 80	38	194	45
80 – 100	7	201	7

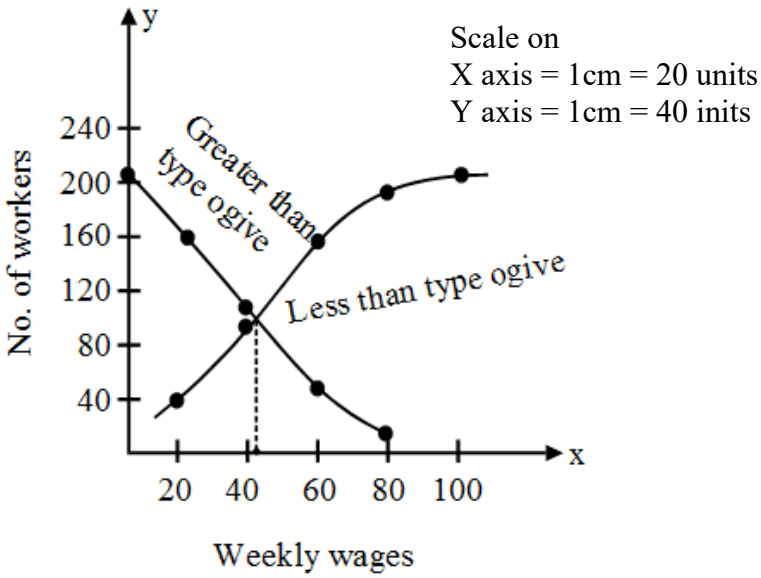
Solution:

Less than curve:

Upper limits of class intervals are marked on the x-axis and less than type cumulative frequencies are taken on y-axis. For drawing less than type curve, points (20, 41), (40, 92), (60, 156), (80, 194), (100, 201) are plotted on the graph paper and these are joined by free hand to obtain the less than ogive.

Greater than ogive:

Lower limits of class interval are marked on x-axis and greater than type cumulative frequencies are taken on y-axis. For drawing greater than type curve, points (0, 201), (20, 160), (40, 109), (60, 45) and (80, 7) are plotted on the graph paper and these are joined by free hand to obtain the greater than type ogive. From the point of intersection of these curves a perpendicular line on x-axis is drawn. The point at which this line meets x-axis determines the median. Here the median is 42.652.



5.5 MODE-ADVANTAGES, AND DISADVANTAGES OF MODE VALUE:

Mode: Mode is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed. If the frequency distribution is regular, then mode is determined by the value corresponding to maximum frequency. There may be situation where concentration observations around some other value. The mode is a poor measure of central tendency when most frequently occurring values of an observation do not appear close to the centre of the data. The mode need not even be a unique value. In such cases, we find the mode for the set of given data. There may or may not be a modal value for a given set of data. For data without any repeating values, there might be no mode at all. Also, we can find data with only one mode, two modes, three modes, or multiple modes. This depends on the given data set.

A list can be unimodal, bimodal, trimodal, or multimodal, depending upon the number of modes it has.

Mode Formula of Ungrouped Data:

Mode for ungrouped data is found by selecting the most frequent item on the list. Now, for any given data range. To find the mode for **ungrouped data**, it just requires the data values to be arranged either in ascending or descending order, then finding the repeated values and their frequency. The observation with the highest frequency is the modal value for the given data is here referred to as the modal value.

Example 8: Compute mode of the following data: 12, 3, 5, 6, 12, 15, 19, 20, 24, 12, 16, 18, 12, 7, 12, 7, 9, 1.

Solution: The number 12 repeated maximum number of times, which is 5 times out of 18 observations. Therefore mode is 12.

Mode Formula of Grouped Data:

Find the maximum class frequency, find the class corresponding to this frequency. It is called the modal class, find the class size. (upper limit – lower limit.)Calculate mode using the formula.

Mode = $l_1 + \frac{(l_2 - l_1)(f_1 - f_0)}{(f_1 - f_0) + (f_1 - f_2)}$, but $(f_1 - f_0) + (f_1 - f_2) = 2f_1 - f_0 - f_2$

Where l = the lower limit of modal class.

h =$(l_2 - l_1)$ the size of class interval.

f_1 = the frequency of the modal class.

f_0 denotes the frequency of the class preceding the modal class.

f_2 denotes the frequency of the class succeeding the modal class.

Example 8: Find the mode of the given data.

Class	50-55	55-60	60-65	65-70
Frequency	2	7	8	4

Solution: Here 8 is the maximum class frequency,

here modal class is 60-65 Class size h = 65-60 = 5, $f_1 = 8$, $f_0 = 7$, $f_2 = 4$, $l_1 = 60$, $l_2 = 61$

Mode =$60 + \frac{(65-60)(8-7)}{(8-7)+(8-4)} = 60 + \frac{(5)(1)}{(1)+(4)} = 60 + \frac{(5)}{(5)}$

Mode = 61

5.6 GRAPHICAL METHOD FOR CALCULATING MODE VALUE.

The procedure of calculating mode using the graphical method is summarized below:

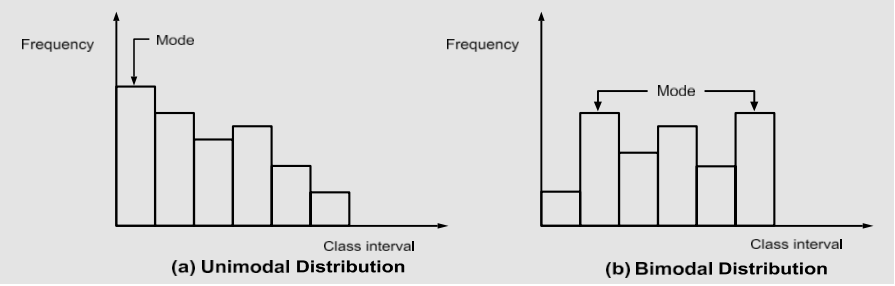
a. Draw a histogram of the data, the tallest rectangle will represent the modal class.

b. Draw two diagonal lines from the top right corner and left corner of the tallest rectangle to the top right corner and left corner of the adjacent rectangles.

c. Draw a perpendicular line from the point of intersection of the two diagonal lines on the x- axis. The value on the x-axis marked by the line will represent the modal value.

Unimodal List: A list of given data with only one mode is called a unimodal list.

Bimodal List**:** A list of given data with two modes is called a bimodal list.



Calculate the mode using the graphical method for the following distribution of data:

Sales (in units) :	53–56	57–60	61–64	65–68	69–72	73–76
Number of days :	2	4	5	4	4	1

Solution: Since the largest frequency corresponds to the class interval 61–64, it is the mode class. Then we have

Mode = $l_1 + \frac{(l_2 - l_1)(f_1 - f_0)}{(f_1 - f_0) + (f_1 - f_2)}$,

here modal class is 61-64 Class size h = 61-64 = 3, $f_1 = 5$, $f_0 = 4$, $f_2 = 4$, $l_1 = 61$

Mode $= 61 + \frac{(64-61)(5-4)}{(5-4)+(5-4)} = 61 + \frac{(3)(1)}{(1)+(1)} = 1 + \frac{(3)}{(2))}$

Mode $= 61 + 1.5 = 62.5$.

Hence, the modal sale is of 62.5 units.

Construct a histogram of the data shown in figure and draw other lines for the calculation of mode value.

The mode value from figure is 62.5 which are same as calculated.

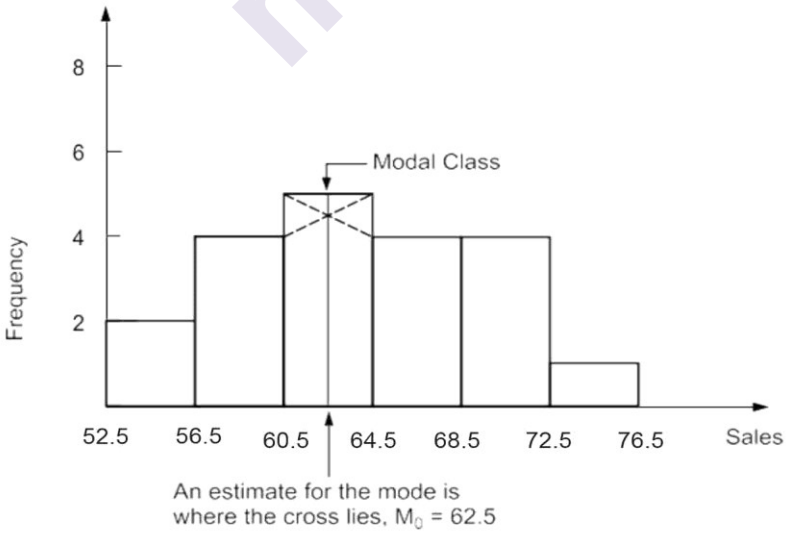


Figure 3.4 Graphs for Modal Value

Advantages and Disadvantages of Mode Value

Advantages:

(i) Mode value is easy to understand and to calculate. Mode class can also be located by inspection.

(ii) The mode is not affected by the extreme values in the distribution. The mode value can also be calculated for open-ended frequency distributions.

(iii) The mode can be used to describe quantitative as well as qualitative data. For example, its value is used for comparing consumer preferences for various types of products, say cigarettes, soaps, toothpastes, or other products.

Disadvantages:

(i) Mode is not a rigidly defined measure as there are several methods for calculating its value.

(ii) It is difficult to locate modal class in the case of multi-modal frequency distributions.

(iii) Mode is not suitable for algebraic manipulations.

(iv) When data sets contain more than one mode, such values are difficult to interpret and compare.

5.7 COMPARISON BETWEEN MEASURES OF CENTRAL TENDENCY

We have already presented three methods to understand the characteristics of a data set. However, the choice of which method to use for describing a distribution of values of observations in a data set is not always easy. The choice to use any one of these three is mainly guided by their characteristics. The characteristics of these three differ from each other with regard to three factors:

Presence of outlier data values Shape of the frequency distribution of data values Status of theoretical development

The Presence of Outlier Data Values: The data values that differ in a big way from the other values in a data set are known as outliers (either very small or very high values). As mentioned earlier, the median is not sensitive to outlier values because its value depend only on the

Number of observations and the value always lies in the middle of the ordered set of values, whereas mean, which is calculated using all data values is sensitive to the outlier values in a data set. Obviously, smaller the number of observations in a data set, greater the influence of any outliers on the mean. The median is said to be resistant to the presence of outlier data values, but the mean is not.

Shape of Frequency Distribution: The effect of the shape of frequency distribution on mean, median, and mode. In general, the median is preferred to the mean as a way of measuring location for single peaked, skewed distributions. One of the reasons is that it satisfies the criterion that the sum of absolute difference (i.e., absolute error of judgment) of median from values in the data set is minimum, that is, □ | x – Med | = min. In other words, the smallest sum of the absolute errors is associated with the median value is the data set as compared to either mean or mode. When data is multi-modal, there is no single measure of central location and the mode can vary dramatically from one sample to another, particularly when dealing with small samples.

The Status of Theoretical Development: Although the three measures of central tendency— Mean, Median, and Mode, satisfy different mathematical criteria but the objective of any statistical analysis in inferential statistics is always to minimize the sum of squared deviations (errors) taken from these measures to every value in the data set. The criterion of the sum of squared deviations is also called least squares criterion. Since A.M. satisfies the least squares criterion, it is mathematically consistent with several techniques of statistical inference.

As with the median, it cannot be used to develop theoretical concepts and models and so is only used for basic descriptive purposes.

5.8 CHAPTER END EXERCISE:

1. On a university campus 200 teachers are asked to express their views on how they feel about the performance of their Union's president. The views are classified into the following categories: Disapprove strongly = 94, Disapprove = 52, Approve= 43, Approve strongly= 11. What is the median view?

2. The following are the profit figures earned by 50 companies in the country:

Profit (in Rs. lakh)	Number of Companies
10 or less	4
20 or less	10
30 or less	30
40 or less	40
50 or less	47
60 or less	50

Calculate

(a) The median, and

(b) The range of profit earned by the middle 80 per cent of the companies. Also verify your results by graphical method.

3. A number of particular items has been classified according to their weights. After trying for two weeks the same items have again been weighted and similarly classified. It is known that the median weight in the first weighting was 20.83 g, while in the second weighing it was 17.35 g. Some frequencies, a and b, in the first weighing and x and y in the second weighing are missing. It is known that a = x/3 and b = y/2. Find out the values of the missing frequencies.

Class	*Frequencies*		*Class*	*Frequencies*	
	I	*II*	*I*		*II*
0–5	*a*	*x*	15–20	52	50
5–10	*b*	*y*	20–25	75	30
10–15	11	40	25–30	22	28

4. The length of time taken by each of 18 workers to complete a specific job was observed to be the following:

Time (in min):	5–9	10–14	15–19	20–24	25–29
Number of Workers :	3	8	4	2	1

I. Calculate the median time

II. Calculate Q1 and Q3

5. The following distribution is with regard to weight (in g) of mangoes of a given variety. If mangoes less than 443 g in weight be considered unsuitable for the foreign market, what is the percentage of total yield suitable for it? Assume the given frequency distribution to be typical of the variety.

Weight (In Kg)	*Number of Mangoes*
410–419	10
420–429	20
430–439	42
440–449	54
450–459	45
460–469	18
470–479	7

Draw an Ogive of 'more than' type of the above data and deduce how many mangoes will be more than 443g.

6. Given the following frequency distribution with some missing frequencies:

Class	*Frequencies*
10-20	185
20-30	--
30-40	34
40-50	180
50-60	136
60-70	--
70-80	50

If the total frequency is 685 and median is 42.6, find out the missing frequencies.

7. What are the advantages and disadvantages of the three common averages: Mean, Median, and Mode?

8. What is a statistical average? What are the desirable properties for an average to possess? Mention the different types of averages and state why arithmetic mean is most commonly used amongst them.

9. The table below is the frequency distribution of ages to the nearest birthday for a random sample of 50 employees in a large company.

Age to nearest birthday	Number of employees
20-29	5
30-39	12
40-49	13
50-59	8
60-69	12

Compute the mean, median and mode for these data.

10. The number of cars sold by each of the 10 car dealers during a particular month, arranged in ascending order, is 12, 14, 17, 20, 20, 20, 22, 22, 24, 25. Considering this scale to be the statistical population of interest, determine the mean, median, and mode for the number of cars sold.

    i. Which value calculated above best describes the 'typical' sales volume per dealer?

For the given data, determine the values at the (i) quartile Q_1 and (ii) percentile P_{30} for these sales amounts

Answers:

1. Disapprove
2. $\text{Med} = 27.5; P_{90} - P_{10} = 47.14 - 11.67 = 35.47$
3. $a = 3, b = 6; x = 9, y = 12$
4. (a) 13.25 (b) $Q_3 = 17.6$, $Q_1 = 10.4$
5. 54.08%; 106
6. 20–30(77)

5.9 SUMMARY:

- **Median:** A measure of central location such that one half of the observations in the data set is less than or equal to the given value.
- **Quartiles:** The values which divide an ordered data set into 4 equal parts. The 2nd quartile is the median.
- **Deciles:** The values which divide an ordered data set into 10 equal parts. The 5th decile is the median.
- **Percentiles:** The values which divide an ordered data set into 100 equal parts. The 50th percentile is the median.
- **Mode value:** A measure of location recognized by the location of the most frequently occurring value of a set of data.
- **Outlier:** A very small or very large value in the data set

5.10 REFERENCES:

1. Fundamentals of, Business Statistics, J. K. Sharma, Pearson Publication.
2. Business Statistics Problems and Solutions, J. K. Sharma, Pearson Publication.
3. Outline of Business Statistics, Kazmier L.J., Schaum's Outline Series, Publisher: McGraw-Hill.

❄❄❄❄❄❄❄

6

OVERVIEW OF MEASURES OF DISPERSION

Unit Structure

6.0 OBJECTIVES

After going through this unit, you will be able to:

- Provide the importance of the concept of dispersion.
- Measure the spread or dispersion, understand it and identify its causes to provide a basis for action.

6.1 INTRODUCTION

We have discussed in our previous unit about average and its various measures of Central Tendency. An average is a single representative numerical figure which summarizes a given data. It tends to be somewhere at the center of a given distribution. However, an average does not tell an entire story regarding the given set of data. This necessity for further description of the data brings us to the concept of dispersion. Dispersion indicates how the data values are scattered away from the average value. This concept brings out how two different distributions with the same average might have a difference in the spread or scattering of values around the average value.

6.2 SIGNIFICANCE OF MEASURING DISPERSION

Measures of dispersion are calculated to help with the following purposes:

1. To review the consistency of measures of central tendency.
2. To compare two or more data about their variability.
3. To control the variability itself.
4. To facilitate the use of another statistical measure.

6.3 CLASSIFICATION OF MEASURE OF DISPERSION

The commonly used measures of dispersion are as follows:

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation

Of the above measures, the first two are positional measures i.e., their values depend on the positional values. The third and fourth measures are called algebraic measures i.e., their values must be calculated.

These dispersion measures are classified into two main categories in terms of the original units of the data or as a figure like a ratio or a percentage. They are:

1. Absolute Measure of Dispersion
2. Relative Measure of Dispersion

The Absolute Measure is a measure of dispersion that is expressed in terms of units. These measures are not used for comparing the data distribution with two different units of measurement.

The Relative Measure is a measure of dispersion expressed in terms of a ratio or a percentage and is independent of units. These measures are commonly used for comparison purposes.

Under Absolute Measures we have:

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation
5. Variance

Under Relative Measures we have:

1. Coefficient of Range
2. Coefficient of Quartile Deviation
3. Coefficient of Mean Deviation
4. Coefficient of Standard Deviation
5. Coefficient of Variance

6.4 CHARACTERISTICS OF A GOOD MEASURE OF DISPERSION

Following are the characteristics which tell whether a measure of dispersion is good or not:

1. It should be simple to understand.
2. It should be easy to calculate.
3. It should be rigidly defined.
4. It should be based on all the observations.
5. It should have sampling stability.
6. It should be useful for further calculations.
7. It should not be affected by extreme values.

6.5 RANGE

It is the simplest measure of dispersion. The Range is defined as the difference between the two extreme observations that is the greatest (maximum) and the smallest (minimum) of the given data.

6.5.1 Absolute measure of range :

(a) For discrete data or raw data:

Range = Highest value of observation – The lowest value of observations.

$= X_n - X_1$

Where X_n is the highest value and X_1 is the lowest value of data.

(b) For grouped and continuous frequency distribution data:

Range = upper limit of the last class –the lower limit of the first class

$= U_l - L_f$

where U_l is the upper limit of the last class and L_f is the lower limit of the first class.

(c) For grouped and discontinuous frequency distribution data:

Range = class mark of last class – the class mark of the first class

$= M_l - M_f$

Where M_l is the class mark of the last class and M_f is the class mark of the first class.

6.5.2 Coefficient of Range (Relative measure of range) :

When two distributions with different units are to be compared, the relative measure that is the coefficient of the range is used.

(a) For discrete data or raw data:

$$\text{Coefficient of Range} = \frac{Highest\ value - Lowest\ value}{Highest\ value + Lowest\ value} = \frac{X_n - X_1}{X_n + X_1}$$

(b) For grouped and continuous frequency distribution data:

$$\text{Coefficient of Range} = \frac{Upper\ limit\ of\ the\ last\ class - Lower\ limit\ of\ the\ first\ class}{Upper\ limit\ of\ the\ last\ class + Lower\ limit\ of\ the\ first\ class} = \frac{U_l - L_f}{U_l + L_f}$$

(c) For grouped and discontinuous frequency distribution data:

$$\text{Coefficient of Range} = \frac{M_l - M_f}{M_l + M_f}$$

Example 1:

Find the range and coefficient of range of the weights in Kg of 8 students.

73, 80, 65, 63, 67, 75, 40, 48

Solution: Here, Highest Value = $X_n = 80$

Lowest Value = $X_1 = 40$

$$\therefore Range = X_n - X_1$$

$$= 80 - 40$$

$$= 40kg$$

$$\text{Coefficient of Range} = \frac{X_n - X_1}{X_n + X_1} = \frac{80 - 40}{80 + 40} = \frac{40}{120} = 0.33$$

Example 2:

The weight of 35 students is given below. Find the range and coefficient of range.

Weight in Kg	30	40	50	60
No. of students	4	5	16	8

Solution:

Here, the highest value in weight $= X_n = 60kg$

The lowest value in weight $= X_1 = 30kg$

$\therefore Range = X_n - X_1$

$= 60 - 30$

$= 30kg$

Coefficient of Range$= \frac{X_n - X_1}{X_n + X_1} = \frac{60 - 30}{60 + 30} = \frac{30}{90} = 0.33$

Example 3:

The yield of grapes in tons from 50 plots is given below. Find the range and coefficient of range.

Yield in tons	0 – 20	20 – 40	40 – 60	60 – 80
No. of Plots	14	10	20	6

Solution:

Here, the given distribution is a grouped and continuous frequency distribution

The upper limit of the last class $= U_l = 80tons$

The lower limit of the first class $= L_f = 0\ tons$

$\therefore Range = U_l - L_f$

$= 80 - 0$

$= 80\ tons$

Coefficient of Range $= \frac{U_l - L_f}{U_l + L_f} = \frac{80 - 0}{80 + 0} = \frac{80}{80} = 1$

Example 4:

The Monthly income distribution of 50 employees is given below. Calculate the range and coefficient of range.

Income in Rs.'000	0 – 2	3 – 5	6 – 8	9 – 11
No. of Employee	5	18	42	27

Solution:

Here, the given distribution is a grouped and discontinuous frequency distribution

Class mark of the last class $= M_l = 10$

Class mark of the first class $= M_f = 1$

$\therefore Range = M_l - M_f$

$= 10 - 1$

$= 9(Rs.'000)$

Coefficient of Range $= \frac{M_l - M_f}{M_l + M_f} = \frac{10-1}{10+1} = \frac{9}{11} = 0.81$

6.5.3. Merits and Demerits of Range:

The following are the merits and demerits of Range as a measure of dispersion:

Merits:

1. Range is easy to calculate.
2. It is easy to understand.

Demerits:

1. It is not based on all the observations in the data.
2. It cannot be used for further mathematical calculations.
3. It is very much affected by fluctuations of sampling.
4. It cannot calculate frequency distribution with open-end classes.
5. It is affected by extreme values.

6.6 QUARTILE DEVIATION

While calculating the Range, we only consider the extreme values of the observation which failed to account for the dispersion within the range. To overcome this difficulty, the Quartile deviation was developed which concentrates on the middle values of the observation.

Quartile Deviation is a measure of dispersion based on the 3rd Quartile (Q_3) and the 1st Quartile (Q_1) of an observation. The inter-quartile range is the difference between the 1st and the 3rd quartile. It describes the extent to which the middle 50% of the observation is scattered or dispersed. It is defined as

$$Inter\ quartile\ range = Q_3 - Q_1$$

Quartile Deviation (Q.D.) is defined as the average difference between the 3rd and 1st quartile. It is also called the semi-inter-quartile range because it represents the half of the inter-quartile range.

$$Quartile\ Deviation\ (Q.D.) = \frac{Q_3 - Q_1}{2}$$

Q.D. as represented above is the absolute measure of dispersion. If it is to be used for comparison study of two observations, relative measure of Q.D. called as the Coefficient of Quartile Deviation is considered. The Coefficient of Q.D. is defined as

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

Example 1:

Find the value of Quartile deviation and coefficient of quartile deviation for the following data:

20, 25, 40, 12, 31, 15, 60

Solution:

To find Q.D., we first calculate $Q_3\ and\ Q_1$.

For which we arrange the data in ascending order, i.e.,

12, 15, 20, 25, 31, 40, 60

Here no. of observations, N = 7

Now, First Quartile Q_1,

$$\begin{aligned} Q_1 &= value\ of\ \left(\frac{N+1}{4}\right)^{th} observation \\ &= value\ of\ \left(\frac{7+1}{4}\right)^{th} observation \\ &= value\ of\ (2)^{nd} observation \\ &= 15 \end{aligned}$$

Similarly, Third Quartile Q_3,

$$\begin{aligned} Q_3 &= value\ of\ 3\left(\frac{N+1}{4}\right)^{th} observation \\ &= value\ of\ 3\left(\frac{7+1}{4}\right)^{th} observation \\ &= value\ of\ (3 \times 2)^{nd} observation \\ &= value\ of\ (6)^{th} observation \\ &= 40 \end{aligned}$$

Hence,

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

$$\text{Quartile deviation} = \frac{40 - 15}{2}$$

$$= 12.5$$

$$Coefficient\ of\ Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40 - 15}{40 + 15} = \frac{25}{55} = 0.455$$

Example: 2

From the following data, calculate the quartile deviation and its coefficient.

X	13	27	38	44
Frequency	10	12	20	18

Solution: We compute the cumulative frequency less than type first,

X	Frequency	Cumulative frequency
13	10	10
27	12	22
38	20	42
44	18	60

$Here, N = \Sigma f = 60$

Q_1

$=$ value of X whose cumulative frequency is just greater than $\left(\frac{N+1}{4}\right)$

$=$ value of X whose cumulative frequency is just greater than $\left(\frac{60+1}{4}\right)$

$=$ value of X whose cumulative frequency is just greater than (15.25)

$= 27$

Similarly,

Q_3

$=$ value of X whose cumulative frequency is just greater than $3\left(\frac{N+1}{4}\right)$

$=$ value of X whose cumulative frequency is just greater than $3\left(\frac{60+1}{4}\right)$

$=$ value of X whose cumulative frequency is just greater than (3×15.25)

$=$ value of X whose cumulative frequency is just greater than (45.75)

$= 44$

Hence,

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

$$\text{Quartile deviation} = \frac{44 - 27}{2}$$

$$= 8.5$$

$$Coefficient\ of\ Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{44 - 27}{44 + 27} = \frac{17}{71} = 0.2394$$

Example 3:

Calculate the semi-inter Quartile Range and its Coefficient from the following data:

Marks	0 – 10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of students	11	18	25	28	30	33	22	15	22

Solution: We compute the cumulative frequency less than type first,

Marks	No. of students	Cumulative frequency
0 – 10	11	11
10 – 20	18	29
20 – 30	25	54
30 – 40	28	82
40 – 50	30	112
50 – 60	33	145
60 – 70	22	167
70 – 80	15	182
80 – 90	22	204

$Here, N = \Sigma f = 204$

Since the given distribution is a grouped continuous frequency distribution,

For 1st quartile

$1st$ quartile class whose cumulative frequency is just greater than $\left(\frac{N}{4}\right)$

$1st$ quartile class whose cumulative frequency is just greater than $\left(\frac{204}{4}\right)$

$1st$ quartile class whose cumulative frequency is just greater than (51)

$\therefore 20 - 30$ *is the* $1st$ *quartile class*.

Here, $l = 20$,

$N = 204$,

$c.f.: Cumulative\ frequency\ of previous\ class\ to\ 20 - 30 = 29$

$f: frequency\ of\ the\ class\ 20 - 30 = 25$

$h: height\ of\ class\ 20-30=10$

$$Q_1 = l + \left(\frac{\left(\frac{N}{4}\right) - c.f.}{f}\right) \times h$$
$$= 20 + \left(\frac{\left(\frac{204}{4}\right) - 29}{25}\right) \times 10$$
$$= 20 + \left(\frac{51-29}{25}\right) \times 10$$
$$= 28.8$$

For 3rd quartile

3rd quartile class whose cumulative frequency is just greater than $3\left(\frac{N}{4}\right)$

3rd quartile class whose cumulative frequency is just greater than $3\left(\frac{204}{4}\right)$

3rd quartile class whose cumulative frequency is just greater than (153)

$\therefore 60-70\ is\ the$ 3rd $quartile\ class.$

Here, $l = 60$,

$N = 204$,

$c.f.: Cumulative\ frequency\ of previous\ class\ to\ 60-70=145$

$f: frequency\ of\ the\ class\ 60-70=22$

$h: height\ of\ class\ 60-70=10$

$$Q_3 = l + \left(\frac{3\left(\frac{N}{4}\right) - c.f.}{f}\right) \times h$$
$$= 60 + \left(\frac{3\left(\frac{204}{4}\right) - 145}{22}\right) \times 10$$
$$= 20 + \left(\frac{153-145}{22}\right) \times 10$$
$$= 63.64$$

Hence,

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2} = \frac{63.64 - 28.8}{2} = 17.42\ marks$$

$$Coefficient\ of\ Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{63.64 - 28.8}{63.64 + 28.8} = 0.38.$$

Note: If the data given is a grouped and discontinuous frequency distribution, convert the data into a continuous frequency distribution and solve according to example 3.

6.6.1 Merits and Demerits of Quartile Deviation:

Merits:

1. It is easy to understand and simple to calculate.
2. It is the only measure of dispersion which can be obtained for open end classes too.
3. It is not affected by extreme values of the observations.

Demerits:

1. It is not based on all the observations as it only concentrates of the middle 50% of the observation.
2. It is not suitable for further mathematical treatment.
3. It is affected by sampling fluctuations.

6.7 MEAN DEVIATION

The range and quartile deviation are positional measures which does not depend on all observations. The mean deviation is a measure of dispersion that is based on all observation. It also measures the amount by which these values in an observation deviate from a measure of central tendency.

The mean deviation is the arithmetic mean of the deviations of the individual values from the average of the given observation. Any of the three averages i.e., mean, median or mode is used for computation. The absolute values of the deviations are used to ignore either the positive or negative signs of the deviations.

Mean deviation is denoted by Greek letter , small delta ∂.

$\delta_{\bar{x}}$ represents the Mean deviation from Mean

δ_{Md} represents the Mean deviation from Median

δ_{M_0} represents the Mean deviation from Mode.

6.7.1 Absolute Measure of Mean Deviation

Mean Deviation is an absolute measure of dispersion.

(a) For discrete or raw data:

If $x_1, x_2, \cdots, x_n$ are the n observations, then

Mean Deviation,

$$M.D. = \frac{\sum_{i=1}^{n} |x_i - A|}{n}$$

where x_i=individual values, n = no. of observations, A = averages: Mean, median or mode

(b) For Frequency distribution:

If $x_1, x_2, \cdots, x_n$ are the n observations and $f_1, f, \cdots, f_n$ are their corresponding frequencies, then

Mean Deviation,

$$M.D. = \frac{\sum_{i=1}^{n} f_i |x_i - A|}{\sum f_i}$$

where A = averages: Mean, median or mode

(c) For grouped frequency distribution:

If $x_1, x_2, \cdots, x_n$ are the n class marks (midpoint) of the class intervals and $f_1, f, \cdots, f_n$ are their corresponding frequencies, then

Mean Deviation,

$$M.D. = \frac{\sum_{i=1}^{n} f_i |x_i - A|}{\sum f_i}$$

where A = averages: Mean, median or mode

6.7.2 Relative Measure of Mean Deviation (Coefficient of Mean Deviation)

The Mean Deviation when divided by the average used for calculating it, we get the Relative Measure of Mean Deviation called as the Coefficient of Mean Deviation. The Coefficient of Mean Deviation is defined as

$$Coefficient\ of\ M.D. = \frac{Mean\ Deviation}{Mean\ or\ Median\ or\ Mode}$$

If the Coefficient of M.D. is desired in percentage, then

$$Coefficient\ of\ M.D. = \frac{Mean\ Deviation}{Mean\ or\ Median\ or\ Mode} \times 100$$

Example 1:

The number of patients seen in the emergency ward of a hospital for a sample of 5 days was 153, 147, 151, 156 and 153. Determine the mean absolute deviation and its coefficient.

Solution: Here number of patients, n=5

The mean of the following data,

$$\bar{x} = \frac{153 + 147 + 151 + 156 + 153}{5} = 152$$

For calculating the mean deviation, we first calculate the absolute deviations.

No. of patients (x_i)	$x_i - \bar{x}$	$\lvert x_i - \bar{x} \rvert$
153	153 – 152 = 1	1
147	147 – 152 = -5	5
151	151 – 152 = -1	1
156	156 – 152 = 4	4
153	153 – 152 = 1	1
Total		$\sum \lvert x_i - \bar{x} \rvert = 12$

$$M.D. = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{12}{5} = 2.4 \cong 3 \; patients \; approx$$

$$Coefficient \; of \; M.D. = \frac{Mean \; Deviation}{Mean} = \frac{2.4}{152} = 0.0158$$

Example 2:

Calculate Mean deviation and its coefficient from median for the following data. Compare the variability.

Year	Sales(Rs. Thousands)	
	Product A	Product B
2006	23	36
2007	41	39
2008	29	36
2009	53	31
2010	38	47

Solution: The median sales of the two products A and B is 38 and 36 respectively.

Product A		Product B	
Sales (x_i)	$\lvert x_i - 38 \rvert$	Sales (x_i)	$\lvert x_i - 36 \rvert$
23	15	31	5
29	9	36	0
38	0	36	0
41	3	39	3

53	15	47	11
n=5	$\sum\|x_i - Md\| = 42$	n=5	$\sum\|x_i - Md\| = 19$

For Product A:

$$M.D. = \frac{\sum_{i=1}^{n}|x_i - Md|}{n} = \frac{42}{5} = 8.4$$

$$Coefficient\ of\ M.D. = \frac{Mean\ Deviation}{Median} = \frac{8.4}{38} \times 100 = 22.1\%$$

For Product B:

$$M.D. = \frac{\sum_{i=1}^{n}|x_i - Md|}{n} = \frac{19}{5} = 3.8$$

$$Coefficient\ of\ M.D. = \frac{Mean\ Deviation}{Median} = \frac{3.8}{36} \times 100 = 10.6\%$$

Since coefficient of M.D. for product A is more than that of product B, we interpretate that sales of product A has greater variability than sales of product B.

6.7.3 Merits and Demerits of Mean deviation

Merits

1. The calculation of Mean Deviation is easy to understand.
2. Unlike Range and Quartile deviation, it is based upon all the observations of the data and shows the dispersion values around the measure of central tendency.
3. It is less affected by extreme values.
4. The problem of average deviation being zero is taken care by considering the absolute value of it.

Demerits

1. It ignores the positive and negative signs of deviations which creates a demand for more reliable measure of dispersion.
2. It is not suitable for further mathematical treatment.
3. Mean deviations give accurate results only when deviations are taken from median. But median does not provide a satisfactory result in case the amount of variation is more in a data set.

6.8 VARIANCE AND STANDARD DEVIATION

In computing the mean deviation, the signs of the deviations are ignored which makes it less reliable as a measure of dispersion. Another way to ignore the signs is to square such values. The sum of all such squared deviations is then divided by the number of observations. This defines

another measure of dispersion which is more convenient for further mathematical treatment is called Variance denoted by σ^2 and defined as

For a set of n observations $x_1, x_2, \cdots, x_n$ with mean $\bar{x}$,

$$Variance, \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

In case the data is grouped in the form of frequency distribution, the individual values of the variable or midpoint of the class-intervals are multiplied with its corresponding frequencies.

For a set of N observations $x_1, x_2, \cdots, x_n$ with its corresponding frequencies $f_1, f, \cdots, f_n$ with mean $\bar{x}$,

$$Variance, \sigma^2 = \frac{1}{\sum f_i}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2$$

6.8.1 Standard Deviation

The variance is difficult to interpret because it is expressed in square units. To measure the dispersion expressed in the units of original data, a positive square root of the variance, called Standard Deviation is taken. The Standard deviation denoted by σ is defined as,

(a) For Raw or ungrouped data:

For a set of n observations $x_1, x_2, \cdots, x_n$ with mean $\bar{x}$,

$$Standard\ Deviation, \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Using the Assumed Mean Method when the arithmetic mean is a fraction value,

$$Standard\ Deviation, \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

Where, $d = x - A$ and A is the assumed mean

(b) For Grouped data:

For a set of N observations $x_1, x_2, \cdots, x_n$ with its corresponding frequencies $f_1, f, \cdots, f_n$ with mean $\bar{x}$,

$$Standard\ Deviation, \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{\sum f_i}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}$$

Using the Assumed Mean Method when the arithmetic mean is a fraction value,

$$Standard\ Deviation, \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum fd^2}{\sum f_i} - \left(\frac{\sum fd}{\sum f_i}\right)^2}$$

Where, $d = x - A$ and A is the assumed mean

Example 1:

Calculate standard deviation from the following set of observations:

8 , 9 , 15, 23, 5, 11, 19, 8, 10, 12

Solution:

The given set of data is an ungrouped set, n = 10

The arithmetic mean $\bar{x} = \frac{8+9+15+23+5+11+19+8+10+12}{10} = 12$

Values x_i	Deviation $(x_i - \bar{x})$	$(x_i - \bar{x})^2$
8	-4	16
9	-3	9
15	3	9
23	11	121
5	-7	49
11	-1	1
19	7	49
8	-4	16
10	-2	4
12	0	0
$\sum x = 120$	$\sum (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 274$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma = \sqrt{\frac{274}{10}} = \sqrt{27.4} = 5.23$$

Example 2:

Calculate the standard deviation from the following data:

Size of item	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

Solution:

The given data is a frequency distribution:

Size of item (x)	Frequency (f)	(fx)	Deviation $(x_i - 9)$	$(x_i - 9)^2$	$f(x_i - 9)^2$
6	3	18	-3	9	27
7	6	42	-2	4	24
8	9	72	-1	1	9
9	13	117	0	0	0
10	8	80	1	1	8
11	5	55	2	4	20
12	4	48	3	9	36
Total	48	432			124

Arithmetic mean,

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{432}{48} = 9$$

$$Standard\ Deviation, \sigma = \sqrt{\frac{1}{\sum f_i} \sum_{i=1}^{n} f_i (x_i - \bar{x})^2}$$

$$\sigma = \sqrt{\frac{124}{48}} = 1.6$$

Example 3:

The following table gives number of finished articles turned out per day by different number of workers in a factory. Find the mean value and the standard deviation of the daily output of finished articles. Also find the variance.

No. of articles (X)	No. of workers	No. of articles (X)	No. of workers

18	3	23	17
19	7	24	13
20	11	25	8
21	14	26	5
22	18	27	4

Solution:

No. of articles	No. of workers	Deviation $d = x_i - A$	fd	d^2	fd^2
18	3	-4	-12	16	48
19	7	-3	-21	9	63
20	11	-2	-22	4	44
21	14	-1	-14	1	14
22 – A	18	0	0	0	0
23	17	1	17	1	17
24	13	2	26	4	52
25	8	3	24	9	72
26	5	4	20	16	80
27	4	5	20	25	100
Total	N=100		38		490

Mean value of the finished articles,

$$\bar{x} = A + \frac{\sum fd}{N} = 22 + \frac{38}{100} = 22.38\ articles$$

Standard deviation,

$$Standard\ Deviation, \sigma = \sqrt{\frac{\sum fd^2}{\sum f_i} - \left(\frac{\sum fd}{\sum f_i}\right)^2}$$

$$Standard\ Deviation, \sigma = \sqrt{\frac{490}{100} - \left(\frac{38}{100}\right)^2} = \sqrt{4.9 - 0.144} = \sqrt{4.756}$$
$$= 2.2\ articles$$

$$Variance = \sigma^2 = (2.2)^2 = 4.84$$

6.8.2. Properties of Standard Deviation

1. The value of S.D. remains the same if in a series each of the observation is increased or decreased by a constant quality.
2. For a given series, if each observation, is multiplied or divided by a constant quality. S.D. will also be similarly affected.
3. For a given set of observations, S.D. is never less than Mean Deviation, i.e., $S.D. \geq M.D.$
4. The standard deviation of first n natural numbers is given by

$$\sigma = \sqrt{\frac{1}{12}(n^2 - 1)}$$

6.8.3 Combined Standard Deviation

The combined standard deviation, σ_{12} of two sets of data containing n_1 and n_2 observations with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviation σ_1 and σ_2, respectively, is given by

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

Where $d_1 = \bar{x}_{12} - \bar{x}_1$; $d_2 = \bar{x}_{12} - \bar{x}_2$ and $\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{\bar{x}_1 + \bar{x}_2}$ (combined arithmetic mean)

Example 1:

For a group of 50 male workers, the mean and standard deviation of their daily wages are Rs. 63 and Rs.9 respectively. For a group of 40 female workers these are Rs. 54 and Rs.6 respectively. Find the standard deviation for the combined group of 90 workers.

Solution:

Given data,

For Male group,

$n_1 = 50, \bar{x}_1 = 63, \sigma_1 = 9$

For Female Group,

$n_2 = 40, \bar{x}_2 = 54, \sigma_2 = 6$

$n_1 + n_2 = 90, \bar{x}_{12} = ?, \sigma_{12} = ?$

The combined formula for arithmetic mean $\bar{x}_{12}$,

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{\bar{x}_1 + \bar{x}_2} = \frac{50 \times 63 + 40 \times 54}{50 + 40} = \frac{5310}{90} = 59$$

$$\therefore d_1 = \bar{x}_{12} - \bar{x}_1 = 59 - 63 = -4$$

$$d_2 = \bar{x}_{12} - \bar{x}_2 = 59 - 54 = 5$$

The Combined standard deviation σ_{12},

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

$$\sigma_{12} = \sqrt{\frac{50((9)^2 + (-4)^2) + 40((6)^2 + (5)^2)}{90}}$$

$$\sigma_{12} = \sqrt{\frac{4850 + 2440}{90}}$$

$$\sigma_{12} = \sqrt{\frac{7290}{90}}$$

$$\sigma_{12} = \sqrt{81} = 9.$$

6.8.4. Merits and Demerits of Standard Deviation

Merits

1. Standard deviation is by far the most important and widely used measure of dispersion. It is rigidly defined and based on all the observations.
2. The squaring of the deviations removes the drawback of ignoring signs of deviation making it more reliable and suitable for further mathematical treatment.
3. It is affected least by the sampling fluctuations.
4. Standard deviation enables to determine the reliability of the means of two or more different series when these means are same. A small S.D. means more clustered and less variability of the values from the mean.
5. It is possible to calculate the combined standard deviation of two or more sets of data.

Demerits

1. It is neither easy to understand nor simple to calculate.
2. It is also affected by the extreme values.

6.9 CHEBYSHEV'S THEOREM

Standard deviation measures the variation among observations in a set of data. If the standard deviation value is small, then values in the data set cluster close to the mean. Conversely, a large standard deviation value indicates that the values are scattered more widely around the mean. The Chebyshev's theorem allows to determine the proportion of data values that fall within a specified number of standard deviation from the mean value.

The theorem states that: For any set of data (population or sample) and any constant z greater than 1 (but need not be an integer), the proportion of the values that lie within z standard deviations on either side of the mean is at least $\{1 - (1/z^2)\}$. That is RF $[\ |x - \mu| \leq z\ \sigma] \geq 1 - 1/z^2$ where RF = relative frequency of a distribution.

Chebyshev's theorem states at least what percentage of values will fall within z standard deviations in any distribution. The relationships involving the mean, standard deviation and the set of observations are called the empirical rule, or normal rule.

Empirical Rule:

For symmetrical, bell-shaped frequency distribution (also called normal curve), the range within which a given percentage of values of the distribution are likely to fall within a specified number of standard deviations (σ)of the mean (μ) is determined as follows:

$\mu \pm \sigma$ covers approximately 68.27 per cent of values in the data set

$\mu \pm 2\sigma$ covers approximately 95.45 per cent of values in the data set

$\mu \pm 3\sigma$ covers approximately 99.73 per cent of values in the data set

6.10 COEFFICIENT OF VARIATION

Standard deviation is the absolute measure of dispersion. The relative measure of dispersion based on standard deviation is called coefficient of standard deviation and is given by

$$Coefficient\ of\ S.D. = \frac{Standard\ deviation}{Mean} = \frac{\sigma}{\bar{x}}$$

100 times the coefficient of standard deviation is called the coefficient of variation (C.V.)

$$C.V. = \frac{Standard\ deviation}{Mean} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

Coefficient of Variation is a pure number independent of units of measurements which makes it very useful for

(i) Comparing two or more data sets expressed in different units of measurements

(ii) Comparing data sets that are in same unit of measurements, but the mean values of data sets are not same.

The lower the value of C.V. indicates uniformity among the values in any data set.

Example 1:

Following are the scores made by two batsmen in one over.

Balls	1	2	3	4	5	6
Batsman A	1	6	4	6	3	1
Batsman B	3	4	4	3	4	3

And $\sigma_1 = 2.062$ for batsman A and $\sigma_2 = 0.5$ for batsman B.

Which of the two batsmen is a better scorer on an average? Which of them is more consistent?

Solution:

Among the two batsmen, the one who has higher arithmetic mean, is the better scorer. And he who has less coefficient of variation, is more consistent in batting.

For Batsman A,

The arithmetic mean,

$$\bar{x}_1 = \frac{\sum x}{n} = \frac{21}{6} = 3.5$$

The Coefficient of Variation,

$$C.V. = \frac{\sigma_1}{\bar{x}_1} \times 100 = \frac{2.062}{3.5} \times 100 = 58.91$$

For Batsman B,

The arithmetic mean,

$$\bar{x}_2 = \frac{\sum x}{n} = \frac{21}{6} = 3.5$$

The Coefficient of Variation,

$$C.V. = \frac{\sigma_2}{\bar{x}_2} \times 100 = \frac{0.5}{3.5} \times 100 = 14.28$$

Since the arithmetic mean for both the batsmen are equal, both are equally good at scoring in terms of average. But since the coefficient of variation for the batsman B is less than batsman A, it is interpreted that batsman B is more consistent in scoring than batsman A.

Example 2:

Find out the coefficient of variation if

i. Standard deviation = 3.5

Number of items = 10

Summation of size of items = 145

ii. Variance = 148.6, Mean = 40

Solution:

i. Given,

$\sigma = 3.5, N = 10, \sum x = 145$

$$\therefore Mean, \bar{x} = \frac{\sum x}{N} = \frac{145}{10} = 14.5$$

Coefficient of Variation,

$$C.V = \frac{\sigma}{\bar{x}} \times 100 = \frac{3.5}{14.5} \times 100 = 24.14$$

ii. Given,

$Variance, \sigma^2 = 148.6, Mean = 40$

$$standard\ deviation = \sqrt{variance} = \sqrt{148.6} = 12.19$$

$$C.V. = \frac{Standard\ deviation}{Mean} \times 100 = \frac{12.19}{40} \times 100 = 30.47$$

6.11 LET'S SUM UP

Dispersion indicates how the data values are scattered away from the average value. The measures used for calculating the spreading of data is called as measures of dispersion. There are various measures of dispersion of which are the following:

i. Range: The difference between the two extreme observations that is the greatest (maximum) and the smallest (minimum) of the given data.

ii. Quartile Deviation: Quartile Deviation (Q.D.) is defined as the average difference between the 3rd and 1st quartile. It is also called the semi-inter-quartile range.

iii. Mean Deviation: The mean deviation is the arithmetic mean of the deviations of the individual values from the average of the given observation.

The best of all the measures,

iv. Standard Deviation: The positive square-root of the variance.

6.12 REFERENCE

J.K. Sharma, Business Statistics, Pearson Education India, 2012.

6.13 CHAPTER END EXERCISE

1. What are the requisites of a good measure of variation?

2. Distinguish between absolute and relative measures of variation. Give a broad classification of the measures of variation.

3. The following are the prices of shares of a company from Monday to Saturday:

Days	Price (Rs.)	Days	Price (Rs.)
Monday	200	Thursday	160
Tuesday	210	Friday	220
Wednesday	208	Saturday	250

Calculate the range and its coefficient.

[Answers: Rs.90, Rs. 0.219]

4. The days sales figures (in Rs) for the last 15 days at Nirula's ice-cream counter, arranged in ascending order of magnitude, is recorded as follows: 2000, 2000, 2500, 2500, 2500, 3500, 4000, 5300, 9000, 12,500, 13,500, 24,500, 27,100, 30,900, and 41,000. Determine the range and middle 50 per cent range for this sample data.

[Ans: Rs. 39,000, Rs. 22,00]

5. The following distribution shows the sales of the fifty largest companies for a recent year:

Sales (Millions of rupees)	Number of Companies
0 – 9	18
10 – 19	19
20 – 29	6
30 – 39	2
40 – 49	5

Calculate the coefficient of range

[Ans: 0.82]

6. For a certain data, the range is 10 and coefficient of range is 0.20, find the smallest and largest values of the data.

[Ans: S = 20, L = 30]

7. Calculate the semi-inter quartile range and its coefficient

Age in years	20	30	40	50	60	70	80
No. of members	3	61	132	153	140	51	3

[Ans: Q.D. = 10, Coefficient = 0.2]

8. From the following data, construct a frequency distribution table starting with 10 – 20 (exclusive method) and calculate the median and the quartile deviation.

   50, 27, 23, 32, 49, 19, 50, 38, 37, 25, 30, 29, 42, 37, 18, 12, 13, 9, 18, 27, 50, 32, 47, 48, 29, 30, 32, 37, 27, 41.

9. You are given the data pertaining to kilowatt hours of electricity consumed by 100 persons in a city.

Consumption (Kilowatt hour)	No. of Users
0 – 10	6
10 – 20	25
20 – 30	36
30 – 40	20
40 – 50	13

   Calculate the quartile deviation and its coefficient from the above-mentioned data

   [Ans: 8.2, 0.32]

10. The following sample shows the weekly number of road accidents in a city during a two-year period:

Number of Accidents	Frequency	Number of Accidents	Frequency
0 – 4	5	25 – 29	9
5 – 9	12	30 – 34	4
10 – 14	32	35 – 39	3
15 – 19	27	40 – 44	1
20 – 24	11		

    Find the interquartile range and coefficient of quartile deviation.

    [Ans: 9.5, 0.3035]

11. The cholera cases reported in different hospitals of a city in a rainy season are given below: Calculate the quartile deviation for the given distribution and comment upon the meaning of your result.

Age Group (Years)	Frequency	Age Group (Years)	Frequency
Less than 1	15	25 – 35	132
1 – 5	113	35 – 45	65
5 – 10	122	45 – 65	46
10 – 15	91	65 and above	15
15 – 25	229		

[Ans: 10.315]

12. You are given the frequency distribution of 292 workers of a factory according to their average weekly income.

Weekly Income (Rs)	No. of Workers	Weekly Income (Rs)	No. of Workers
Below 1350	8	1450 – 1470	22
1350 – 1370	16	1470 – 1490	15
1370 – 1390	39	1490 – 1510	15
1390 – 1410	58	1510 – 1530	9
1410 – 1430	60	1530 and above	10
1430 – 1450	40		

Calculate the quartile deviation and its coefficient from the above-mentioned data.

[Ans: 27.78, 0.020]

13. Calculate the mean deviation from mean and median for the following data:

100, 150, 200, 250, 360, 490, 500, 600, 671

Also calculate coefficient of mean deviation.

[Ans: M.D.(Mean) = 174.44, M.D.(Median) = 173.3, Coefficient of M.D. = 0.48]

14. Define mean deviation. State the merits and demerits.

15. Calculate the mean deviation from (i) arithmetic mean, (ii) mode and (iii) median in respect of the marks obtained by nine students: 7, 4, 10, 9, 15, 12, 7, 9, 7

[Ans: (i) 2.35, (ii) 2.55 (iii) 2.33

16. With median as base, calculate mean deviation and compare the variability of the two series A and B:

A	B	A	B
2354	50010	3020	70110
5000	100000	3541	83001
2780	61061	5150	91100
3011	70005		

[Ans: M.D.(A) = 79.29, Coefficient = 0.26

M.D.(B) = 13277.85, Coefficient = 0.189, A is more variable than B]

17. Find the average deviation from mean for the following distribution:

Quantity demanded (in units) : 60 61 62 63 64 65 66 67 68

Frequency : 62 60 15 29 25 12 10 64 63

[Ans: M.D. = 2.83]

18. Find the average deviation from mean for the following distribution:

Dividend yield	0 – 3	3 – 6	6 – 9	9 – 12	12 – 15	15 – 18	18 – 21
No. of Companies	2	7	10	12	9	6	4

[Ans: 3.823]

19. Find the average deviation from median for the following distribution:

Sales (Rs.'000)	1 – 3	3 – 5	5 – 7	7 – 9	9 – 11	11 – 13	13 – 15	15 – 17
No. of shops	6	53	85	56	21	26	4	4

[Ans: Md = 6.612, avg deviation = 2.30]

20. Following are the monthly sales of a firm in a year:

Months	1	2	3	4	5	6	7	8	9	10	11	12
Sales	50	30	25	44	48	51	55	60	42	35	28	40

Find out the mean deviation from mean for the above data.

[Ans: 2.76]

21. Explain with suitable examples the term dispersion. Mention some common measures of dispersion and describe the one which you think to be most important of them.

22. What are the different measures of dispersion? How would you calculate them for a given frequency distribution ? Briefly discuss the relative merits of different measures of dispersion.

23. The following data give the annual crop in kilograms from 10 experimental farms:

    2020 2100 2040 2030 2070 2060 2080 2050 2110 2090

    Find the mean and its standard deviation of the annual yield.

    [Ans: 2065, 28.72]

24. The annual salaries of a group of employees are given in the following table:

Salaries (in Rs.'000)	45	50	55	60	65	70	75	80
No of Persons	3	5	8	7	9	7	4	7

    Calculate the mean and standard deviation of the salaries.

    [Ans: Rs.63600, Rs.10344]

25. Find the standard deviation and mean of breaking strength of 80 test pieces of certain alloy from the following table:

Breaking strength	Number of pieces
44 – 46	3
46 – 48	24
48 – 50	27
50 – 52	21
52 – 54	5

    [Ans: Mean = 49.025, S.D. = 1.949]

26. Blood serum cholesterol levels of 10 persons are as under: 240 260 290 245 255 288 272 263 277 250. Calculate the standard deviation with the help of assumed mean.

    [Ans: 16.48]

27. The means of two samples of size 50 and 100 respectively are 54.1 and 50.3 and the standard deviation are 8 and 7. Find the standard deviation of the combined group of 150 samples.

[Ans: 7.56]

28. The standard deviation of a distribution of 100 values was Rs 2. If the sum of the squares of the actual values was Rs 3,600, what was the mean of this distribution?

[Ans: 5.66]

29. Two salesmen selling the same product show the following results over a long period of time:

	Salesman X	Salesman Y
Avg Sales volume per month (Rs.)	30,000	35,000
Standard Deviation	2,500	3,600

Which salesman seems to be more consistent in the volume of sales?
[Ans: Salesman X]

30. The number of employees, average daily wages per employee, and the variance of daily wages per employee for two factories are given below:

	Factory A	Factory B
No. of employees	50	100
Average daily wages(Rs.)	120	85
Variance of daily wages(Rs)	9	16

In which factory is there greater variation in the distribution of daily wages per employee?

[Ans: Factory B]

31. The share prices of a company in Mumbai and Kolkata markets during the last ten months are recorded below:

Month	Mumbai	Kolkata
January	105	108
February	120	117

March	115	120
April	118	130
May	130	100
June	127	125
July	109	125
August	110	120
September	104	110
October	112	135

Determine the arithmetic mean and standard deviation of prices of shares. In which market are the share prices more stable?

[Ans: Mumbai]

Practical on Measures of Dispersion Using R Programming

1. Monthly sales (in '00 Rs.) of 20 small shops are given below.

   120,115,130,140,180,210,180,120,130,150,100,190,210,160,150,160, 190,200,170,152

   Calculate range, coefficient of range, Q1, Q3, quartile deviation, coefficient of quartile deviation, mean deviation about mean, variance, standard deviation and coefficient of variation.

   Code:

```
x=c(120,115,130,140,180,210,180,120,130,150,100,190,210,160,150,160,190,200,170,152)
x
n=length(x)
n
am=mean(x) # Mean
am
min=min(x)
max=max(x)
range=max-min # Range
range
cr=r/(max+min) #Coefficient of Range
cr
q1=quartile(x,0.25) # 1st Quartile
q1
```

```
q3=quartile(x,0.75) # 3rd Quartile
q3
qd=(q3-q1)/2 # Quartile deviation
qd
cqd=(q3-q1)/(q3+q1) # Coefficient of Quartile Deviation
cqd
md=sum(abs(x-am))/n # Mean Deviation
md
v1=var(x) # Variance
v1
sd=sd(x) # Standard Deviation
sd
cv=sd*100/am # Coefficient of Variation
cv
```

2. Calculate quartile deviation, coefficient of quartile deviation, standard deviation and coefficient of variation for the following data:

Class Interval	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 - 60	60 – 70	70 – 80
Frequency	4	8	15	24	16	14	12	7

Code:

```
lc=seq(0,70,10)
lc
uc=seq(10,80,10)
uc
h=10
f=c(4,8,15,24,16,14,12,7)
f
N=sum(f)
N
x=(lc+uc)/2
x
am=sum(f*x)/N
am
lcf=cumsum(f)
```

```
lcf
q1c=min(which(lcf>=N/4))
q1c
q1=lc[q1c]+((N/4)-lcf[q1c-1])*h/f[q1c] # 1st Quartile
q1
q3c=min(which(lcf>=3*N/4)) # 3rd Quartile
q3c
q3=lc[q3c]+((3*N/4)-lcf[q3c-1])*h/f[q3c]
q3
qd=(q3-q1)/2 # Quartile deviation
qd
cqd=(q3-q1)/(q3+q1) # Coefficient of Quartile deviation
cqd
v=sum(f*(x-am)^2)/N # Variance
v
sd=v^0.5 # Standard Deviation
sd
cv =sd*100/am # Coefficient of Standard Deviation
cv
```

❄❄❄❄❄❄❄

7

MOMENTS, SKEWNESS AND KURTOSIS

Unit Structure

7.0 OBJECTIVE

After successful completion of this course, learners would be able to

1. Define raw moments and central moments.
2. Obtain relationship between raw moments and central moments.
3. Apply Sheppard's correction for moments.
4. Identify & examine the shape of distributions.
5. Study the Skewness and Kurtosis of data.

7.1 INTRODUCTION

The Central Tendency and Dispersion measures are insufficient to represent the data distribution. They may have the same central tendency and dispersion despite their variances in nature and composition. We can discriminate between different shapes of frequency distributions using Skewness & Kurtosis. To make a proper comparison between two or more distributions, we have to study four characteristics of the distributions namely average, variation, skewness, and kurtosis. Skewness and Kurtosis are concerned with the shape of a distribution. But for studying Skewness and Kurtosis one must be familiar with the concepts of moments.

Moments for a distribution:

Definition:

Moments can be defined as arithmetic mean of various powers of deviation of the variable taken from an arbitrary point (number) A.

(Usually A is Mean, Median, Mode or Zero.)

Moments about arbitrary point:

Let a variable X takes values $x_1, x_2, x_3, \ldots, x_n$. Then r^{th} moment about point 'A' is denoted by $'\mu_r'$ and is defined as

$$\mu_r = \frac{\sum_{i=1}^{n}(x_i - A)^r}{n}, \text{ Where } r = 1, 2, 3\ldots$$

Let a variable X takes values $x_1, x_2, x_3, \ldots, x_n$. Then first moment about point 'A' is denoted by $'\mu_1'$ and defined as

$$\mu_1 = \frac{\sum_{i=1}^{n}(x_i - A)}{n}$$

Similarly, we can define μ_2, μ_3, μ_4 (called the second, the third and the fourth moments respectively).

For the above individual distribution or series moments are defined as:

$\mu_2 = \frac{\sum_{i=1}^{n}(x_i - A)^2}{n}$ (Called the second moment about point 'A')

$\mu_3 = \frac{\sum_{i=1}^{n}(x_i - A)^3}{n}$ (Called the third moment about point 'A')

$\mu_4 = \frac{\sum_{i=1}^{n}(x_i - A)^4}{n}$ (Called the fourth moment about point 'A')

If the variable X takes values $x_1, x_2, x_3, \ldots, x_n$ with corresponding frequencies $f_1, f_2, f_3, \ldots, f_n$, then r^{th} moment about point 'A' is denoted by $'\mu_r'$ and is defined as

$\mu_r = \frac{\sum_{i=1}^{n} f_i(x_i - A)^r}{N}$, Where r = 1, 2, 3… & N = $\sum_{i=1}^{n} f_i$

Note: For the continuous grouped frequency distribution, x_i can be taken as the class-mark of i_{th} class interval and

$\mu_r = \frac{\sum_{i=1}^{n} f_i(x_i - A)^r}{N}$, can be used to determine the r^{th} moment about point 'A'.

Types of Moments

There are two types of moments

a) Moments about Zero or Origin (Raw Moments):

b) Moments about Mean (Central Moments):

1. **Raw moments:** Raw moment is defined as the arithmetic mean of various powers of deviation of the variable taken from **zero**. (Here A is zero, i.e. A = 0)

For Ungrouped Data:-

Let a variable X takes values $x_1, x_2, x_3, \ldots, x_n$. Then r^{th} raw moment is denoted by μ'_r and is defined as

$$\mu'_r = \frac{\sum_{i=1}^{n}(x_i)^r}{n}, \text{ Where r = 1, 2, 3…}$$

For Grouped Data:-

If the variable x takes values $x_1, x_2, x_3, \ldots, x_n$ with corresponding frequencies $f_1, f_2, f_3, \ldots, f_n$, then r^{th} raw moment is denoted by μ'_r and is defined as

$$\mu'_r = \frac{\sum_{i=1}^{n} f_i(x_i)^r}{N}, \text{ Where r = 1, 2, 3… \& N} = \sum_{i=1}^{n} f_i$$

Note: If the arbitrary point A is other than the mean of distribution then $\mu'_1, \mu'_2, \mu'_3 \text{ and } \mu'_4$ are respectively called the first, the second, the third and the fourth **raw moments or non-central moments.**

2. **Central Moments:** Central moment is defined as the arithmetic mean of various powers of deviation of the variable taken from **arithmetic mean**. (Here A is $\bar{x}$, i.e. A = $\bar{x}$)

For Ungrouped Data:-

Let a variable X takes values $x_1, x_2, x_3, \ldots, x_n$. Then r^{th} central moment is denoted by $'\mu_r'$ and is defined as

$$\mu_r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^r}{n}, \text{ Where r = 1, 2, 3…}$$

For Grouped Data:-

If the variable X takes values $x_1, x_2, x_3, \ldots, x_n$ with corresponding frequencies $f_1, f_2, f_3, \ldots, f_n$, then r^{th} central moment is denoted by $'\mu_r'$ and is defined as

$$\mu_r = \frac{\sum_{i=1}^{n} f_i(x_i-\bar{x})^r}{N}, \text{ Where r = 1, 2, 3… \& N} = \sum_{i=1}^{n} f_i$$

Relation between Raw and Central Moments:

Let a variable X takes values $x_1, x_2, x_3, \ldots, x_n$.

Then r_{th} raw moment of X is given by $\boldsymbol{\mu'_r = \frac{\sum_{i=1}^{n}(x_i)^r}{n}}$, Where r = 1, 2, 3…

And r_{th} central moment of X is given by $\boldsymbol{\mu_r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^r}{n}}$, Where r = 1, 2, 3…

1. Let r = 1

First raw moment is given by $\boldsymbol{\mu'_1 = \frac{\sum_{i=1}^{n}(x_i)}{n} = \bar{x}}$

First central moment is given by

$$\boldsymbol{\mu_1 = \frac{\sum_{i=1}^{n}(x-\bar{x})}{n}}$$

$$= \frac{\sum_{i=1}^{n} x - \sum_{i=1}^{n} \bar{x}}{n}$$

$$= \frac{\sum_{i=1}^{n} x}{n} - \frac{\bar{x}\sum_{i=1}^{n} 1}{n}$$

$$= \bar{x} - \bar{x} = 0$$

$$\mu_1 = 0$$

2. Let r = 2

Second raw moment of X is given by $\mu'_2 = \frac{\sum_{i=1}^{n}(x_i)^2}{n}$

Second central moment of X is given by

$$\mu_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$= \frac{\sum_{i=1}^{n}((x_i)^2 - 2x_i\bar{x} + \bar{x}^2)}{n}$$

$$= \frac{\sum_{i=1}^{n}(x_i)^2}{n} - \frac{2\bar{x}\sum_{i=1}^{n} x_i}{n} + \frac{\sum_{i=1}^{n}\bar{x}^2}{n}$$

$$= \mu'_2 - 2{\mu'_1}^2 + \mu'_1$$

$$\mu_2 = \mu'_2 - {\mu'_1}^2$$

Also it can be shown that

3. $\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2{\mu'_1}^3$

4. $\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2{\mu'_1}^2 - 3{\mu'_1}^4$

Karl Pearson's Beta (β) and Gamma (γ) coefficients based on moments

Karl Pearson's has given the following coefficients based on the first four central moments.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3};\ \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3};\ \gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3$$

All these coefficients are pure numbers being free from units of measurements and hence these are useful for comparative studies. These are used as a measure of skewness and kurtosis.

Note: another coefficient based on moments known as **alpha (α) coefficient** is also sometimes used. The four alpha coefficients are:

$$\alpha_1 = \frac{\mu_1}{\sigma} = 0;\ \alpha_2 = \frac{\mu_2}{\sigma^2} = 1\ (\mu_2 = \sigma^2)$$

$$\alpha_3 = \frac{\mu_3}{\sigma^3} = \sqrt{\beta_1} = \gamma_1;$$

$$\alpha_4 = \frac{\mu_4}{\sigma^4} = \beta_2$$

Sheppard's Correction for Moments:

In case of grouped frequency distributions, while calculating moments, the frequencies of the class interval are taken to be the frequencies of the corresponding mid-values of the class intervals. This is based on the assumptions that the frequencies are concentrated at the mid-points of the corresponding classes. Although this assumptions is approximately correct is case of symmetrical or moderately asymmetrical distributions yet it is not true in general and introduces some errors knows as 'grouping error' in the estimation of moments.

W.F. Sheppard's has shown that for grouped frequency distribution, if the frequency tappers off to zero in both directions, the effect due to grouping at the mid-points of the intervals can be corrected by the following formula knows as "**Sheppard's Correction".**

$$\mu_2\,(Corrected) = \mu_2\,(Calculated) - \frac{h^2}{12}$$

$$\mu_3(Corrected) = \mu_3\,(Calculated)$$

$$\mu_4(Corrected) = \mu_4\,(Calculated) - \frac{1}{2}h^2\mu_2 + \frac{7}{240}\,h^4$$

Where, h is width of class interval (uniform). The above corrections are valid only for symmetrical or slightly symmetrical distributions.

Note: These corrections should be preferably applied only if the total frequency is fairly large.

Solved Problems:

Q.1) The first two moments of a distribution about the value 1 are 3 and 16 respectively. Find the mean of the distribution.

Solution: From the given problem, we have

$$A = 1,\ \mu'_1 = 3,\ \mu'_2 = 16$$

To Find:

1) The mean of the distribution

$$\mu'_1 = \frac{\sum_{i=1}^{n}(x_i - A)}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} A}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} A}{n}$$

$$\mu_1 = \bar{x} - A$$

$$\bar{x} = \mu'_1 + A = 3 + 1 = 4$$

Therefore, the mean of distribution is 4.

Q.2) Calculate the first four moments of the following distribution about the mean and hence find β_1 and β_2.

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

Solution:

Set d = x - 4

x	**f**	**d**	**fd**	**fd^2**	**fd^3**	**fd^4**
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
Total	**256**	**0**	**0**	**512**	**0**	**2816**

Moments about the point x = 4 are

$$\mu'_1 = \frac{\sum fd}{N} = 0$$

$$\mu'_2 = \frac{\sum fd^2}{N} = \frac{512}{256} = 2$$

$$\mu'_3 = \frac{\sum fd^3}{N} = 0$$

$$\mu'_4 = \frac{\sum fd^4}{N} = \frac{2816}{256} = 11$$

Moments about mean are:

$\mu_1 = 0$

$\mu_2 = \mu'_2 - \mu'^2_1 = 2$

$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1 = 0$

$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1 = 11$

$\boldsymbol{\beta_1 = \frac{\mu^2_3}{\mu^3_2} = 0 \ , \beta_2 = \frac{\mu_4}{\mu^2_2} = \frac{11}{4} = 2.75}$

7.3 SKEWNESS

Definition: Skewness means 'lack of symmetry'. We study skewness to have an idea about the shape of the curve which can be drawn with the help of given data.

Distribution is said to be skewed if –

1. Mean, median and mode fall at different points.
2. Quartiles are not equidistance from median; and
3. The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to other.

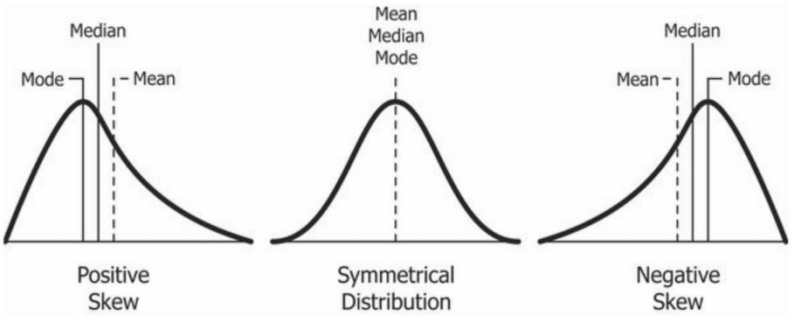
Note:

1. A distribution is said to be **symmetric** about its arithmetic mean (A.M.) if the deviation of the values of the distribution from their A.M. are such that corresponding to each positive deviation, there is negative deviation of the same magnitude.
2. If the distribution is **symmetric** then $\mu_3 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{n} = 0$ & $\mu_3 = \frac{\sum_{i=1}^{n} f_i(x_i-\bar{x})^3}{N} = 0$. For continuous grouped frequency data
3. If a distribution is not symmetric then the distribution is called a skewed distribution.
4. A skewed distribution is also called as asymmetric distribution.
5. Thus in case of a **skewed distribution** the magnitudes of the positive and the negative deviations of the values from their mean do not balance.

Types of Skewness:

Skewness is of two types

1. **Positive Skewness**
2. **Negative Skewness**



(Fig. 01)

1. **Positive Skewness**: Skewness is **positive** if the larger tail of the distribution lies towards the higher values of the variate (the right), i.e. if the curve drawn with the help of the given data is stretched more to the right than left and the distribution is said to be **positively skewed distribution**.

   For a positively asymmetric distribution: **A.M. > Median > Mode**

2. **Negative Skewness**: Skewness is **negative** if the larger tail of the distribution lies towards the lower values of the variate (the left), i.e. if the curve drawn with the help of the given data is stretched more to the left than right and the distribution is said to be **negatively skewed distribution**.

   For a negatively asymmetric distribution: **A.M. < Median < Mode**

Note:

1. For a symmetric distribution: A.M. = Median = Mode
2. Skewness is positive if A.M. > Median or A.M. > Mode
3. Skewness is negative if A.M. < Median or A.M. < Mode

Measure of Skewness

Various measures of skewness are (these are absolute measures of skewness)

1. S_k = Mean – Median
2. S_k = Mean – Mode
3. $Sk = (Q_3 – M_d) – (M_d – Q_1)$

Relation Measure of Skewness

As in dispersion, for comparing two series we do not calculate these absolute measures but we calculate the relative measures called the **coefficient of Skewness,** which is pure numbers **independent of units of measurement**.

The following are the **coefficients of Skewness-**

1. Prof. Karl Pearson's first measure of skewness (S_k) = $\frac{(A.M.-Mode)}{\sigma}$

2. Prof. Karl Pearson's second measure of skewness (S_k) = $\frac{3(A.M.-Median)}{\sigma}$

3. Prof. Bowley's coefficient of skewness (based on quartiles)

$S_k = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$ or $S_k = \frac{Q_3+Q_1-2Median}{Q_3-Q_1}$

4. $S_k = \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3}$

Note:

1.

   a. If the coefficient of skewness is positive (i.e. $S_k > 0$), then the distribution is positively skewed.

   b. If the coefficient of skewness is negative (i.e. $S_k < 0$), then the distribution is negatively skewed.

   c. If the coefficient of skewness is zero (i.e. $S_k = 0$) then the distribution is normal distribution (no skewed).

2.

   a. The distribution is symmetric, if $\sqrt{\beta_1} = 0 \; or \; \mu_3 = 0$.

   b. The distribution is positively skewed, if $\sqrt{\beta_1} > 0 \; or \; \mu_3 > 0$.

   c. The distribution is negatively skewed, if $\sqrt{\beta_1} < 0 \; or \; \mu_3 < 0$.

Remark:

Bowley's coefficient of skewness is also known as **Quartile Coefficient of Skewness** and is especially useful in situations where quartiles and median are used,

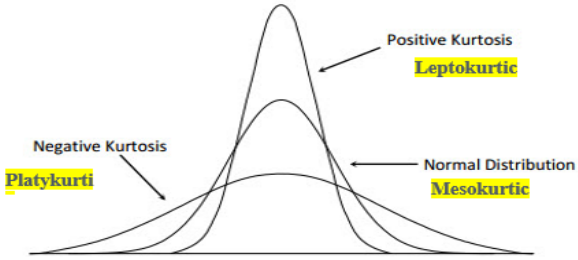
I. When the mode is ill-defined and extreme observations are present in the data.

II. When the distributions has open end classes or unequal class intervals.

In these situations Pearson's coefficients of skewness **cannot** be used.

7.4 KURTOSIS

The three measures namely, measures of central tendency, measure of variations (moments) and measure of skewness that we have studied so far are not sufficient to describe completely the characteristics of a frequency distribution. Neither of these measures is concerned with the peakedness of a frequency distribution.



Kurtosis is concerned with the flatness or peakedness of frequency curve – The graphical representation of frequency distribution.

Definition:

Clark and Schkade defined kurtosis as: "Kurtosis is the property of a distribution which express its relative peakedness."

Types of Kurtosis:

1. **Mesokurtic**
2. **Leptokurtic**
3. **Platykurtic**

1. **Mesokurtic:** The frequency curve which is bell shaped curve is considered as standard and such distribution is called Mesokurtic.

   The normal curve is termed Mesokurtic.

2. **Leptokurtic:** A curve which is more peaked than the normal curve is called Leptokurtic. For **Leptokurtic curve kurtosis is positive** and dispersion is least among all the three types.

3. **Platykurtic:** A curve which is flatter than the normal curve is called Platykurtic. For **Platykurtic curve kurtosis is negative** and dispersion is more.

Measure of Kurtosis:

Kurtosis is measured by $\boldsymbol{\beta_2}$ ***and*** $\boldsymbol{\gamma_2}$. These are called **Karl Pearson's Coefficient of kurtosis**. These are defined as below:

$$\beta_2 = \frac{\mu_4}{\mu^2{}_2}, or\ \beta_2 = \frac{\mu_4}{\sigma^4}$$

$$\text{And} \qquad \gamma_2 = \beta_2 - 3$$

Percentile Coefficient of Kurtosis K= $\frac{Q}{P_{90} - P_{10}}$

Note:

1.

a. For Mesokurtic curve (normal curve), $\beta_2 = 3 \; or \; \gamma_2 = 0$

b. For a Leptokurtic curve, $\beta_2 > 3 \; or \; \gamma_2 > 0$

c. For a Platykurtic curve, $\beta_2 < 3 \; or \; \gamma_2 < 0$

2. The two terms 'frequency distribution' and 'frequency curve' are synonymous because a frequency curve is graphical representation of a frequency distribution. Thus Mesokurtic cuve relates to mesokurtic distribution, etc.

Solved Problems:

Q.1) Find the skewness and kurtosis for the following distribution by the method of moments:

Number of hours worked	1-3	3-5	5-7	7-9
Number of boys	3	5	1	1

Solution:

For the data we have, $\bar{x} = \frac{\sum fx}{N} = \frac{40}{10} = 4$

Let d = $x - \bar{x}$

X	**f**	**d**	**d^2**	**fd^2**	**fd^3**	**fd^4**
2	3	-2	4	12	-24	48
4	5	0	0	0	0	0
6	1	2	4	4	8	16
8	1	4	16	16	64	256
Total	10			32	48	320

$s^2 = \mu_2 = \frac{\sum fd^2}{N} = \frac{32}{10} = 3.2$

$\mu_3 = \frac{\sum fd^3}{N} = \frac{48}{10} = 4.8$

$\mu_4 = \frac{\sum fd^4}{N} = \frac{320}{10} = 32.0$

Skewness: $\beta_1 = \frac{{\mu^2}_3}{{\mu^3}_2} = \frac{4.8^2}{3.2^3} = 0.0703$

Kurtosis: $\beta_2 = \frac{\mu_4}{{\mu^2}_2} = \frac{32.0}{3.2^2} = 3.125$

Alternatively $\gamma_1 = \sqrt{\beta_1} = \sqrt{0.0703} = 0.265$

$$\gamma_2 = \beta_2 - 3 = 3.125 - 3 = 0.125$$

Hence, $\gamma_1 > 0$, the skewness is positive and $\gamma_2 > 0$, the distribution is leptokurtic.

Q.2) Compute the coefficient of skewness (S_k) if

a. Mean = 125, Mode = 95, S.D. =5

b. Mean = 16, Mode = 16, S.D. = 5

Solution:

a. $S_k = \frac{Mean - Mode}{S.D.} = \frac{125-95}{5} = \frac{30}{5} = 6$

Since, $S_k > 0$ the distribution is positively skewed.

b. $S_k = \frac{Mean - Mode}{S.D.} = \frac{16-16}{5} = \frac{0}{5} = 0$

Since, $S_k = 0$ the distribution is Symmetric.

7.5 CHAPTER END EXERCISE

Q.1) Fill in the blanks.

1. The sum of squares of deviations is least when measured from ----------.
2. The sum of 10 items is 12 and sum of their squares is 16.9. Variance = ----------.
3. In any distribution, the standard deviation is always -------- the mean deviation from mean.
4. In a symmetric distribution, the mean and mode are ----------.
5. If mean, mode and standard deviation of a frequency distribution are 41, 45 and 8 respectively, then its Pearson's coefficient of skewness is ---------.
6. For a symmetric distribution $\beta_1 =$ ---------
7. If $\beta_2 > 3$, the distribution is said to be---------
8. If the mean and the mode of a given distribution are equal, then its coefficient of skewness is ---------.

9. If the kurtosis of a distribution is 3, it is called -------- distribution.

10. In a frequency curve of scores, the mode was found to be higher than the mean, this shows that the distribution is ---------.

Q.2) State whether the following statements are true and false. In each false statement give the correct statement.

1. Mean, standard deviation and variance have the same unit.

2. $\beta_2 \geq 1$ is always satisfied.

3. $\beta_1 = 0$ is a conclusive test for a distribution to be symmetrical.

4. Two distributions have the same values of mean, S.D. and Skewness must have the same Kurtosis.

5. For a Platykurtic distribution, $\gamma_2 > 0$.

Q.3) For the following questions given correct answers.

1. For any frequency distribution, the Kurtosis is

   a. Greater than 1

   b. Less than 1

   c. Equal to 1

   d. NOTA

2. The measure of kurtosis is

   a. $\beta_2 = 0$

   b. $\beta_2 = 3$

   c. $\beta_2 = 4$

   d. $\beta_2 = 1$

3. In a frequency scores, the mode was found to be higher than the mean, this shows the distribution is

   a. Symmetric

   b. Negatively Skewed

   c. Positively Skewed

   d. NOTA

4. The statement that the variance is equal to second central moment is:

   a. Always true

b. Sometimes true

c. Never true

d. Ambiguous

5. The limits for quartiles coefficient of skewness are

a. ± 3

b. 0 and 3

c. ± 1

d. $\pm \infty$

Q.4) Answer the following questions.

1. Define raw and central moments of a frequency distribution.
2. What are 'Skewness' and 'Kurtosis'? Bring out their importance in describing frequency distributions.
3. Find the first four central moments for the data 15, 16, 17, 18, 19, 20.
4. Find the first four moments of the set 2, 3, 7, 8, 10.
5. The first four moments from mean of a distribution are 0, 3.2, 3.6 and 20. The mean value is 11. Calculate the first four moments about zero.
6. For a frequency distribution given below, calculate the coefficient of skewness based on quartiles:

Class Limits	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
Frequency	5	9	14	20	25	15	8	4

7. Compute Karl Pearson's coefficient of skewness for the following distribution:

Class Limits	130-134	135-139	140-144	145-149	150-154	155-159	160-164
Frequency	3	12	21	28	19	12	5

8. Karl Pearson's coefficient of Skewness of a distribution is 0.32, its standard deviation is 6.5 and mean is 29.6. Find the mode of distribution.
9. In a frequency distribution, the coefficient of Skewness based upon the quartiles is 0.6. If the sum of the upper and lower quartiles is 100 and median is 38, find the value of the upper and lower quartiles.

10. Find the second, third and fourth central moments of a frequency distribution given below. Hence find measure of skewness and measure of kurtosis.

Class Limits	110.0 - 114.9	115.0 - 119.9	120.0 - 124.9	125.0 - 129.9	130.0 - 134.9	135.0 - 139.9	140.0 - 144.9
Frequency	5	15	20	35	10	10	5

11. For a given data, find the coefficient of Skewness

    a. Mean = 8, Mode = 8, S.D. = 4

    b. $Q_1 = 2$, $Q_3 = 8$, Median = 5

12. The first four moments of a distribution about the value 4 of the variable are – 1.5, 17, -30 and 108. Find the moment about mean, β_1 and β_2.

13. For a distribution of 250 heights, calculations showed that the mean, standard deviation, $\beta_1$and β_2 were 52 inches, 3 inches, 0 and 3 inches respectively. It was however, discovered on checking that the two items 64 and 50 in the original data were wrongly written in place of the correct values 62 and 52 inches respectively. Calculate the correct frequency constants.

7.6 REFERENCE

I have taken help from the following books in writing this book. I am indebted to the authors of these books.

1. A textbook of Business Statistics – Padmalochan Hazarika
2. Fundamentals of Mathematical Statistics – S.C. Gupta & V. K. Kapoor
3. Mathematics & Statistics for Economics – G.S. Monga
4. Skew diagram taken from source: https://commons.wikimedia.org
5. Kurtosis diagram taken from source: https://www.analyticsvidhya.com

❄❄❄❄❄❄❄