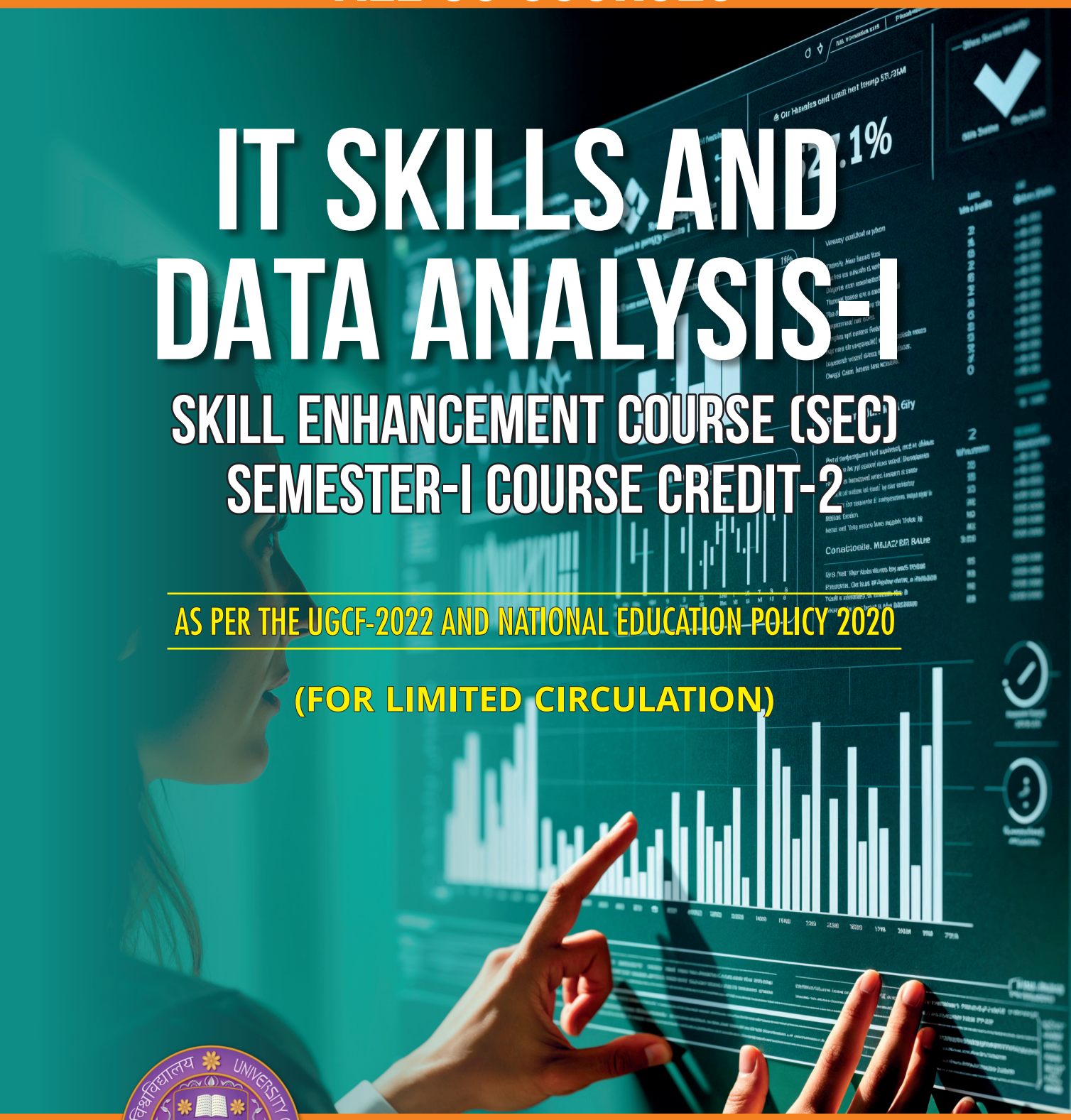
ALL UG COURSES

IT SKILLS AND DATA ANALYSIS-I

SKILL ENHANCEMENT COURSE (SEC)
SEMESTER-I COURSE CREDIT-2

AS PER THE UGCF-2022 AND NATIONAL EDUCATION POLICY 2020

(FOR LIMITED CIRCULATION)



DEPARTMENT OF DISTANCE AND CONTINUING EDUCATION
CAMPUS OF OPEN LEARNING, SCHOOL OF OPEN LEARNING,
UNIVERSITY OF DELHI

[FOR LIMITED CIRCULATION]

Editor

Prof. Shikha Gupta

Content Writers

Dr. Archana Verma, Ms. Priyanka Gupta

Academic Coordinator

Deekshant Awasthi

Department of Distance and Continuing Education

E-mail: ddceprinting@col.du.ac.in
computerscience@col.du.ac.in

Published by:
Department of Distance and Continuing Education
Campus of Open Learning, School of Open Learning,
University of Delhi, Delhi-110007

Printed by:
School of Open Learning, University of Delhi

Reviewer

Ms. Asha Yadav

Corrections/Modifications/Suggestions proposed by Statutory Body, DU/ Stakeholder/s in the Self Learning Material (SLM) will be incorporated in the next edition. However, these corrections/modifications/suggestions will be uploaded on the website https://sol.du.ac.in. Any feedback or suggestions may be sent at the email- feedbackslm@col.du.ac.in

Printed at: **Taxmann Publications Pvt. Ltd., 21/35, West Punjabi Bagh, New Delhi - 110026 (500 Copies, 2024)**

Department of Distance & Continuing Education, Campus of Open Learning, School of Open Learning, University of Delhi

Contents

UNIT - I

LESSON

1

Introduction to Statistics

STRUCTURE

1.1 Introduction

The present-day society is essentially information-oriented. In various fields, we need information. The set of information in the form of numerical figures is known as **data.** The numerical figures maybe for example about: imports and exports of different countries, per-capita national income, minimum and maximum temperature, food production, increase in population, per-capita expenditure, income tax, sales tax and property tax, etc. Thus, Data may be defined as the figures which are numerical or otherwise collected with a definite purpose and from which meaningful information can be obtained. The branch of study which deals with data is known as Statistics.

Statistics is a branch of science which deals with the collecting, organising, summarising, presenting and analysing data and drawing valid conclusions and thereafter making reasonable decisions on the basis of such analysis.

The word **'statistics'** seems to have been derived from the Latin word 'status' or the Italian word 'statista' or the German word 'statistic' each of which means a "political state". In ancient times, the government used to collect the information regarding the "population" and

Notes

"property" of the country. It was confined only to the affairs of the state but now it embraces almost every sphere of human activity. It is now finding wide application in almost all sciences – social as well as physical – such as biology, psychology, education, economics, business management, etc. It has become indispensable in all phases of human endeavor.

1.2 Types of Data

There are two types of statistical data: (i) Primary data and (ii) Secondary data.

Primary data: When an investigator collects the data himself with a definite plan in mind, it is called **primary data.** Primary data is reliable and relevant because they are original in nature and collected by some institution or some individuals.

For example: The investigator goes from house to house to collect the information about number of males and females in a family.

Secondary data: If it is not possible to collect data due to lack of time and resources, the data collected by other investigators or official data or data published in newspapers can be used. The data which is collected by one investigator and used by another for his study is called **secondary data.** Since these data are collected for a certain purpose, and if used for different purpose, the same data may not be relevant for the user. Therefore, such data must be used with great care.

Primary data is the data which is collected for the first time by the investigator and therefore original in character while secondary data is the data which has already been collected by other persons and has passed through the statistical process and procedure machine at least once. Thus, we can say data which is secondary in the hands of one may be primary for other.

1.3 Collection of Data

Methods of the collection of primary data and secondary data would not be exactly identical because in one case the data have to be originally collected while in the other the work is of the nature of compilation. There are various methods of the collection of primary and secondary data and the 'choice of the method depends on a number of factors.

Notes

(i) **Availability of Finance:** This is one of the factors which influences the selection of the method of collection of data. When financial resources at the disposal of the investigator are scanty, he shall have to leave aside expensive methods, even though they are better than others which are comparatively cheap.

(ii) **Availability of Time:** Some methods, involve long duration of enquiry while with others the enquiry can be conducted in a comparatively shorter duration. The time at the disposal of the investigator thus affects the selection of the technique by which data is to be collected.

1.4 Collection of Primary Data

Four types of the collection of primary data:

(i) Direct investigation

(ii) Indirect oral investigation

(iii) Schedules and questionnaires

(iv) Local reports

(i) **Direct Investigation:** The investigator collects the information directly from the concerned sources. He has to be on the spot for conducting the enquiry and has to meet people from whom data have to be collected. It is necessary that in such cases the investigator has a keen sense of observation and he is very polite and courteous. He should further acquaint himself with local conditions, customs and traditions so that he is in a position to identify himself fully the persons from whom the information is sought. In some cases, it may not be possible or worthwhile to contact directly the persons concerned and in such cases the investigator has to cross-examine other persons who are closely in touch with the sources of data. The information elicited in such a manner should be carefully used and the investigator should make sure that the persons from whom data are being collected actually know the facts fully and can deliver him the goods. The investigator has to be very tactful and cautious in such cases. He should put easy and simple questions which are capable of being answered precisely and in a language which is not vague.

Notes

The method of direct personal investigation is suitable only for intensive investigations. It involves enormous cost and usually requires a long time. It is naturally not suitable for extensive enquiries where the scope of investigation is wide. Further, in this method the bias or prejudice of the investigator can do a lot of damage as he is sole in charge of the collection of data. This method, however, gives very satisfactory results if the scope of the enquiry is narrow and if the investigator is fully dependable and is completely unbiased.

(ii) **Indirect Oral Investigation:** When the above-mentioned method cannot be used either on account of the reluctance of persons to part with information when approached directly, or on account of the extensive scope of the enquiry or on account of some other reason an indirect oral examination can be conducted. In this method data are not collected directly from the persons concerned but through indirect sources. Persons who are supposed to have knowledge about the problem under investigation are interrogated and the desired information is collected. Usually in such enquiries a small list of questions relating to the investigation is prepared and these questions are put to different persons (known as witnesses) and their answers are recorded. Most of the commissions and committees appointed by the Government to collect statistical data or to carry on such investigations in which factual data have to be compiled, make use of this method. In this method the accuracy of data collected would largely depend on the type of persons whose evidences are being recorded. It is, therefore, necessary to be very cautious in the selection of these persons.

(iii) **Schedules and Questionnaires:** An important method for the collection of data followed usually by private individuals, research workers, non-official institutions and sometimes the Government also, is that of schedules and questionnaires. In this method a list of questions relating to the problem under investigation is prepared and printed. Thereafter information is collected from various sources in any of the following ways:

(i) ***By sending the questionnaire to the persons concerned and requesting them to answer the questions and return the questionnaire:*** The main advantage of this method is that it is least expensive, and with it information can be collected

Notes

from a wide area in a comparatively short period of time. If the investigation is properly conducted the method can easily ensure a reasonable standard of accuracy. Success in this method depends on the co-operation that the informants are prepared to give. Generally, it has been found that the informants adopt an attitude of indifference towards such enquiries and in many cases do not even return the questionnaire. Even those who answer the questions do so most haphazardly and in a very vague and unintelligible manner. In order to have correct answers the investigator should send a very polite letter to the informants emphasizing the need and usefulness of the investigation that is being conducted and requesting them to give their cooperation by sending correct replies. He should further give them an assurance that if the informants so wish their replies would be kept confidential. Further the questions that are asked should be very carefully framed. The questions should be:

(a) Short and clear

(b) Easy to understand and answer

(c) Few in number

(d) Free from ambiguity

(e) Such as can be answered in Yes or No if opinion is sought on a particular point.

(f) Corroboratory in nature

(g) Not such which call for confidential information

(h) Not such which may hurt the sentiments of the informants or may arouse resentment in their minds.

However, this method cannot be used if the informants are illiterate.

(ii) By sending the questionnaires through enumerators to help the informants in filling the answers: In this method the enumerators go to the informants along with the questionnaires and help them in recording their answers. The enumerators explain the aims and objects of the investigation to the informants and also emphasize the necessity and usefulness of correct answers. They also remove the difficulties which

Notes

any informant may feel in understanding the implications of a particular question or the definition or concept of difficult terms. This method is very useful in extensive enquiries and with it, fairly dependable results can be expected. It is however, very expensive and usually such enquiries can be conducted only by the Government. Population census all over the world is conducted by this method. In such enquiries it is necessary that not only the questions are simple and few in number but the enumerators are also courteous and polite and have proper training.

(iv) **Local Reports:** In this method the data is collected through local agents or correspondents in their own way and to their own likings. This method yields results promptly and easily and are least expensive. However, such data cannot be very reliable and is used in those cases where the purpose of investigation can be served with rough estimates only and where high degree of precision is not necessary.

1.5 Sources of Secondary Data

We know that secondary data are those which have already been collected and analysed by someone else, and as such the problems associated with the original collection of data do not arise here. Secondary data may be either published or unpublished. The sources of published data are usually:

(a) Official publications of the central, state and the local governments.

(b) Official publications of the foreign government or international bodies like the United Nations Organization and its subsidiary bodies.

(c) Reports and publications of trade associations, chambers of commerce, banks, co-operative societies, stock exchanges, and trade unions, etc.

(d) Technical trade journals like the Economics, Indian Journal of Economics, Commerce, Capital, etc., and books and newspapers.

(e) Reports submitted by economists, research scholars, university bureaus and various other educational associations, etc.

These sources of unpublished data are varied, and such materials may be found with scholars and research workers, trade associations, chambers of commerce, labour bureaus, etc. Many enquiries of a private nature are

Notes

conducted by these bodies and these findings are not published and are usually meant for the consumption of their members only.

Editing and Scrutiny of Secondary Data

The secondary data must be used with caution. It is usually very difficult to verify such data and to edit them to find out inconsistencies, probable errors and omissions. Scrutiny of the secondary data is essential because the data might be inaccurate, unsuitable or inadequate, In the words of Bowley, "It is never safe to take published statistics at their face value without knowing their meanings and limitations and it is always necessary to criticise arguments that can be based on them." Statistics collected by other people cannot be fully depended upon as they 'may contain many pitfalls and unless they have been thoroughly scrutinized, they should not be used. The secondary data should possess the following attributes:

(i) They should be reliable. The reliability of the data can be tested by finding out:

(a) Who collected the data and- from which sources?

(b) Are both the compiler and the source dependable?

(c) Were the data collected by the use of proper methods?

(d) At what time were the data collected? Can it be regarded as normal time?

(e) Are there any possibilities of deliberate or unconscious bias on the part of the compiler?

(f) What degree of accuracy was desired by the compiler? Was it achieved?

They should be suitable for the purpose of investigation. Even if the data are reliable, they should not be used if they are found to be unsuitable for the purpose of investigation. Data which are suitable for one enquiry may be entirely unsuitable for another. The definition of various terms and units of collection must also be carefully scrutinized and the object, scope and nature of the enquiry should also be properly studied. If there are differences in these, the data are not fit to be used. They should be adequate. The data may be found to be reliable and suitable but they may be inadequate for the purpose of the enquiry. The original data may refer to an area which is wider or narrower than the area of the present enquiry and if it is so, they should not be used, because there might be

Notes

significant variations in different regions. Further the data may not cover suitable periods; for a monthly study of a phenomenon; yearly figures are inadequate. Again, the degree of accuracy achieved in the data may be found to be inadequate for the purpose of the investigation in which they are proposed to be used. Thus, it is very risky to use statistics collected by other people unless they have been thoroughly scrutinized and found reliable, suitable and adequate.

1.6 Quantitative Data and Qualitative Data

Primary and Secondary data can be further classified as:

Quantitative Data: Quantitative data are counts or can be expressed as numbers. These can be measured, e.g. length, time, temperature, marks, score in cricket match etc.

Qualitative Data: Qualitative data are descriptive and cannot be counted or measured. These data depict qualities or characteristics, e.g., honesty, beauty, intelligence, illness etc.

Discrete Data: If the observations differ from one another by exact magnitude, i.e., it takes integral values then it is called discrete data. This data is countable. For example, number of students in a school, size of shoes, number of runs in a match etc. are discrete data.

Continuous Data: If the values are not of exact magnitude but are of interval type then the data is called continuous data. This data is measurable. If we take two real values, it can take on all real values between them. For example, time, height, weight, length etc. are continuous data.

Time Series Data: It is a collection of observations obtained through repeated measurements over time. Time series data is a sequence of data points indexed in time order or data collected at different points in time. For example, stock price, annual retail sales, weather report, heart rate etc. are time series data.

Geographical Data: If place, location or geographical division is the main factor in the data then we get geographical data. For example, number of cancer patients in different states, number of fields under a particular crop in different districts of a state, etc.

Notes

1.7 Levels of Measurements

Scales of measurement or level of measure describes the nature of information within the numbers assigned to variables. Best known classification is with four levels or scales of measurement are nominal, ordinal, interval, and ratio.

1. **Nominal Scale** of measurement deals with variables that are non-numeric or numbers without any value. That is the measurement is said to be on nominal scale if the observations are taken in accordance with some attribute or quality. For example, married or unmarried; literate or illiterate; male or female; nationality, religion, etc. In nominal scaling the numerical values are categorized in such a way that they are mutually exclusive and collectively exhaustive. Since nominal scale has no order and no arithmetic origin, it is said to be least powerful among four scales.
2. **Ordinal Scales** The scale is said to be ordinal when some definite order or rank is also given along with the nominal scale. For example, if the data is collected on the basis of intelligence i.e. genius, above average, average, dull, etc. and are ranked as 1, 2, 3 and so on. In ordinal scaling the numerical values are categorized to denote qualitative differences among the various categories as well as rank ordered in some meaningful way according to some preference. However, the ordinal scale does not give any indication of the magnitude of difference among the ranks. Therefore, ordinal scale of measurement looks at variables where the order matters but the; differences do not matter. If A is better than B, which is better than C, and so on. But is A four times better than D? Is it two times better? As, the order is important but not the differences.
3. **Interval Scales** In the interval scale the data are represented in a definite interval. Interval scale is interpretable, i.e. it not only classifies individuals according to certain categories and determine order of these categories, it also measures the magnitude of the differences in the preferences among the individuals and we can perform arithmetic operations on the data collected.

   Example is temperature as the difference between each value is the same difference between 45 and 30 degrees is measurable 15

Notes

degrees, also the difference between 90 and 75 degrees. Time is everyday example in which the increments are known, consistent, and measurable of an Interval Scale.

4. **Ratio Scales** of measurement is the most informative scale. It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured. A ratio scale can be considered as the three earlier scales rolled up in one. Like a nominal scale, it provides a name or category for each object, the numbers as labels. Like an Ordinal Scale, the objects are ordered ordering of the numbers. Similar to an interval scale, the same difference at two places on the scale has the same meaning. And in addition, the same ratio at two places on the scale also carries the same meaning.

Ratio scales are the ultimate to measurement scales because they tell us the order, the exact value between units and also have an absolute zero which allows for both descriptive and inferential statistics. Everything about interval data applies to ratio scales to have a clear definition of zero examples include height and weight.

Ratio scales provide possibilities to statistical analysis, as variables can be added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

1.8 Presentation of Data

The data obtained in the original form is called raw data or ungrouped data. This data is condensed into groups or classes in order to study their salient features. Such an arrangement is called presentation of data.

Raw data can be arranged as follows:

(i) Serial order

(ii) Ascending order

(iii) Descending order

If the raw data is arranged in ascending or descending order of magnitude, it is called an array.

Notes

For example, suppose 50 students obtained the following marks in a Mathematics test.

41, 23, 6, 32, 22, 18, 13, 38, 15, 18, 19, 2, 7, 9, 10, 19, 15, 22, 22, 41, 11, 49, 31, 17, 28, 37, 4, 6, 27, 29, 36, 26, 3, 29, 20, 15, 48, 17, 20, 40, 45, 46, 45, 48, 49, 47, 46, 35, 32 and 34.

To get wiser picture, we arrange them in either ascending or descending order. Now we arrange the data given in above example in ascending order.

2, 3, 4, 6, 6, 7, 9, 10, 11, 13, 15, 15, 15, 17,17, 18, 18, 19, 19, 20, 20, 22, 22, 22, 23, 26, 27, 28, 29, 29, 31, 32, 32, 34, 35, 36, 37, 38, 40, 41, 41, 45, 45, 46, 46, 47, 48, 48, 49, 49.

But this method also does not reduce the bulk of the data. The most appropriate measure is to represent the data in tabular form which we will be discussing in the next chapter.

1.9 Exercise

1. Divide your class into five groups and ask them to collect data from day to day life.
2. Classify the data collected in question no. in primary and secondary data.
3. Discuss the meaning and scope of Statistics.
4. What do you understand by the word 'Statistics' in (i) Singular form (ii) Plural form.
5. Define some fundamental characteristics of Statistics.
6. What are primary and secondary data? Which of the two is more reliable and why?
7. Explain the purpose and methods of classification of data.
8. Distinguish between primary and secondary data. What are the various methods used in collecting primary data. Examine the relative merits and limitations of each method.
9. "In collection of statistical data commonsense is the chief requisite and experience the chief teacher." Discuss the above statement with comments.

Notes

10. Mention the different kinds of statistical methods generally used in investigations. Are there any fields of enquiry where these methods cannot be used?

11. "Though figures cannot lie, yet liars can figure". Expand the above statement so as to explain its bearing on the use of secondary statistical data.

12. How will you organise an investigation into the handloom weaving industry of Uttar Pradesh? Prepare a questionnaire for the purpose.

13. How far do the results of statistical investigations depend upon correct sampling? Compare the methods used to secure representative data.

14. State and explain the law of statistical regularity. Discuss the methods generally used in sampling.

15. Compare the different methods used in the collection of statistical data. Explain the importance of determining a statistical unit in the collection of data.

16. Distinguish between a census and a sample enquiry and briefly discuss their comparative advantages. Which of these methods would you prefer for calculating the total wages of workers in a given industry?

LESSON

2

Frequency Distribution

STRUCTURE

2.1 Introduction

Classification of the data helps in organizing raw data into smaller groups which facilitates comparison. It helps in studying the relationship between several characteristics and facilitates further statistical treatment.

Primary rules that should be followed while classifying are:

1. The classes should be unambiguously defined.
2. Every observation must belong to one class or the other i.e. classes should be exhaustive.
3. As far as possible, classes should be of equal width.

The number of classes should neither be too large nor too small.

2.2 Frequency Distribution

The number of times an observation occurs is called the **frequency**.

The tabular arrangement of data showing the frequency of each item is called a frequency distribution.

Notes

Thus, there are two types of frequency distributions, of grouped data.

1. **Inclusive form (Discontinuous form):** A frequency distribution in which both lower and upper limit of each class is included in the class.
2. **Exclusive form (Continuous form):** A frequency distribution in which upper limit of each class is excluded and lower limit is included.

2.3 Frequency Distribution of an Ungrouped Data

Tally method. A bar (|) called tally mark is put against the number when it occurs. When occurred 4 times, the fifth occurrence is represented by putting diagonally across tally (\) on the first four tallies. This technique facilitates the counting of the tally marks at the end.

Example 1: The number of books sold per day at a bookseller shop during the month of June, 2002 are given below:

46, 44, 50, 50, 45, 50, 42, 43, 50, 46, 41, 42, 41, 46, 44, 50, 42, 43, 50, 42, 45, 45, 45, 41, 42, 43, 44, 43, 43, 43

Prepare a frequency distribution table by tally methods.

Solution: In the figure column of the table, we write all the numbers from lowest to highest. In second column we put tally marks and in the third column we write down the frequency.

Frequency Distribution of Books

No. of Books	Tally Marks	Frequency
41	|||	3
42	𝍸||	5
43	𝍸 |	6
44	|||	3
45	||||	4
46	|||	3
50	𝍸 |	6
	Total	**30**

Notes

Example 2: The distance (in km) of 40 female engineers from their residence to their place of work were found as follows:

5	3	10	20	25	11	13	7	12	31
19	10	12	17	18	11	32	17	16	2
7	9	7	8	3	5	12	15	18	3
12	14	2	9	6	15	15	7	6	12

Construct a grouped frequency distribution table with class size 5 for the data given above, taking the first interval as 0-5 (5 not included). What main features do you observe from this tabular representation?

Solution: Frequency distribution of above data in tabular form is as follows:

Distances (in km)	Tally Marks	Frequency
0–5	𝍸	5
5–10	𝍸 𝍸 |	11
10–15	𝍸 𝍸 |	11
15–20	𝍸 ||||	9
20–25	|	1
25–30	|	1
30–35	||	2
Total		**40**

We observe that:

(i) The residence of 22 female engineers is within 5 to15 km.

(ii) The residence of 4 female engineers is within 20 to 35 km.

Example 3: The relative humidity (in%) of a certain city for a month of 30 days was as follows:

98.1	98.6	99.2	90.3	86.5	95.3	92.9	96.3	94.2	95.1
89.2	92.3	97.1	93.5	92.7	95.1	97.2	93.3	95.2	97.3
96.2	92.1	84.9	90.2	95.7	98.3	97.3	96.1	92.1	89

(i) Construct a grouped frequency distribution table with classes 84–86, 86–88 etc.

(ii) Which month or season do you think this data is about?

(iii) What is the range of this data?

Notes

Solution:

(i) Frequency distribution of above data in tabular form as

Relative Humidity (in%)	Frequency
84–86	1
86–88	1
88–90	2
90–92	2
92–94	7
94–96	6
96–98	7
98–100	4
Total	30

(ii) This data is related to the rainy season.

(iii) Range 99.2 – 84.9 = 14.3

Example 4: The following data represents the life times in hour of 20 bulbs produced by a factory:

235, 236, 238, 232, 230, 239, 235, 236, 237, 231, 240, 239, 233, 232, 240, 231, 239, 238, 233, 230.

Construct a frequency table for the given data. Use your table to answer the following questions.

(i) How many bulbs had life time more than 235 hours?

(ii) How many bulbs had lifetime less than 234 hours?

(iii) What percent of bulbs had lifetime more than 238 hours?

(iv) The manufacturer claimed that 70% of the bulbs produced by his factory has a minimum lifetime of 236 hours. What percent of the bulbs did not fulfil the claim?

Solution: Frequency Distribution of bulbs is as follows:

Lifetime (in Hours)	Tally Marks	Frequency
230	||	2
231	||	2
232	||	2
233	||	2

Notes

Lifetime (in Hours)	Tally Marks	Frequency
234	-	0
235	||	2
236	||	2
237	|	1
238	||	2
239	|||	3
240	||	2
Total		**20**

1. Number of bulbs having life time more than 235 hours = 10

2. Number of bulbs having life time less than 234 hours = 8

3. Percentage of bulbs having life time more than 238 hours

$= 5/20 \times 100$

$= 25\%$

4. Percentage of bulbs having minimum life time of 236 hours.

$= 10/20 \times 100$

$= 50\%$

Claim by manufacture is 70%

Actual percentage is 50%

Therefore, percentage of bulbs which did not fulfil the claim will be 70% – 50% = 20%

Example 5: The marks obtained by 30 students in a class in a test out of 10 marks are as follows:

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

Make a frequency distribution table for the above data. Use the table to find:

(i) The number of students passed, if the minimum pass marks is 40%?

(ii) How many students failed?

(iii) How many students secured the highest marks?

(iv) How many students received more than 60% of marks?

Notes

Solution:

Marks Obtained	Tally Marks	Frequency
0	||	2
1	|	1
2	||||	4
3	$\cancel{\text{IIII}}$	5
4	||||	4
5	|||	3
6	|||	3
7	|||	3
8	||	2
9	|||	3
10		0
Total		**30**

(i) Number of students who got pass marks = 18 Ans.

(ii) Number of students failed = 12 Ans.

(iii) Number of students who secured highest marks = 3 Ans.

(iv) Number of students who received more than 60% = 8 Ans.

Sometimes the data is so large that it is inconvenient to list every mark in the frequency distribution table. Then we group the marks into convenient intervals. Consider the following data of the marks obtained by 30 students: 84, 75, 93, 48, 60, 57, 61, 67, 53, 39, 50, 66, 81,49, 54, 88, 78, 64, 35, 69, 46, 52, 73, 5

2.4 Procedure of Arranging the Given Data Into Class Intervals

- **Step 1 - Range:** Determine the difference between the minimum and maximum marks. This is called range of the data.

  Range = Maximum mark – Minimum mark. Here the maximum and minimum marks are 93 and 35 respectively. Therefore, the Range = 93 – 35 = 58.
- **Step 2 - Class Size:** Decide the class size. Let the class size be 10.
- **Step 3 - Number of Classes:** Divide the range into groups such that the number of groups may lie between 5 and 15.

Notes

Let the class size = 10

No. of classes = Range/Class Size

Range = 58 and Class size = 10

Thus, number of classes = 5.8 = 6 (say)

We should have 6 classes, each of size 10.

The 6 classes are 35 – 44, 45 – 54, 55 – 64, 65 – 74, 75 – 84 and 85 – 94. The first class 35 – 44 includes the minimum value 35 and the maximum value 44.

Class Intervals: Each of the groups formed above is called a class intervals or simply a class.

- **Step 4 - Maximum Value and Minimum Value:** The minimum value of the variate should be included in the first class. The maximum value of the variate should be included in the last class interval.
- **Step 5 - Tally:** The marks in the proper class as shown in the table below. After four tally marks put a slant line for the fifth mark cutting the previous four tally marks. Each class - intervals covers 10 marks. This (10) is called the size of the class.

Class Intervals	Tally Marks	Frequency
35 – 44	|||	3
45 – 54	𝍸 ||	7
55 – 64	𝍸	5
65 – 74	𝍸 |	6
75 – 84	𝍸	5
85 – 94	||||	4
Total		30

Class Limits: The lowest and highest marks which are included in a class are called lower class limit and upper-class limit of the class. For example, in the class interval 35 – 44, the lower-class limit is 35 and the upper-class limit is 44. This type of class intervals is called discontinuous or Inclusive form since both 35 and 44 are included in the class interval 35 – 44.

Class Boundaries/Actual Class Limits/True Class Limits: Assume that the marks are given in fractions. Which class will include the marks 44.5?

Notes

It should be included in 35 – 44 or 45 – 54? The class boundary is calculated as the average of the upper- and lower-class limits of the adjacent classes, i.e.,

[44 (upper limit of first class) + 45 (lower limit of second class)]/2 = 89/2 = 44.5

The above table is re-written as:

Class	Actual Class	Frequency
35 – 44	34.5 – 44.5	3
45 – 54	44.5 – 54.5	7
55 – 64	54.5 – 64.5	5
65 – 74	64.5 – 74.5	6
75 – 84	74.5 – 84.5	5
85 – 94	84.5 – 94.5	4
Total		**30**

1. In class 54.5 – 64.5, the lower-class boundary is 54.5 and the upper-class boundary is 64.5.
2. Class size is (64.5 – 54.5) = 10, is same for each class.
3. Marks 54.5 will be included in the class 54.5 – 64.5 while marks 64.5 will be included in the next class i.e., 64.5 – 74.5.

These types of class intervals are called continuous or exclusive form since the upper limit of the class is excluded from the class and is included in the next class.

Class Mark: Class mark is the mid value of a particular class i.e., the average of its class limits or class boundaries.

For example, for the class 35 – 44, the class mark = (35 + 44)/2 = 39.5

Class mark is the representative of its class.

Class Mark (or Mid-Value of the class) = (Upper class limit + lower class limit)/2

Example 6: The following data gives marks out of 60 obtained by 30 students of a class in a test:

50, 22, 56, 47, 27, 37, 40, 16, 12, 33, 29, 49, 35, 15, 43, 29, 31, 22, 51, 27, 29, 27, 22, 18, 20, 11, 19, 31, 23, 58.

Notes

Arrange them in ascending order and present it as a grouped data (i) in Inclusive form and (ii) in Exclusive form.

Solution: Arranging the marks in ascending order we get 11, 12, 15, 16, 18, 19, 20, 22, 22, 22, 23, 27, 27, 27, 29, 29, 29, 31, 31, 33, 35, 37, 40, 43,47,49, 50, 51, 56, 58.

(i) Inclusive Form

Frequency Distribution of Marks

Marks (Class Intervals)	Tally Marks	Frequency
11 – 20	𝍸 ||	7
21 – 30	𝍸 𝍸	10
31 – 40	𝍸 |	6
41 – 50	||||	4
51 – 60	|||	3
Total		**30**

(ii) Exclusive Form

Frequency Distribution of Marks

Marks (Class Intervals)	Tally Marks	No. of Students (Frequency)
11 – 21	𝍸 ||	7
21 – 31	𝍸 𝍸	10
31 – 41	𝍸 |	6
41 – 51	||||	4
51 – 61	|||	3
Total		**30**

From the above example, both the inclusive and exclusive methods give us the same class frequency although the class intervals are apparently different in two cases.

Class interval 11 – 21 means marks obtained are 11 and more but less than 21. Clearly, 21 does not belong to this class 11 – 21.

Example 7: The monthly wages of 30 workers in a factory are given below:

830, 835, 890, 810, 835, 836, 869, 845, 898, 890, 820, 860, 832, 833, 855, 845, 804, 808, 812, 840, 885, 835, 836, 878, 840, 868, 890, 806, 840, 890.

Notes

Represent the data in the form of a frequency distribution with class size 10.

Solution: Minimum monthly wage = 804 Maximum monthly wage = 898

Range = Maximum monthly wage – Minimum monthly wages = 898 – 804 = 94

Size of the class interval = 10

Range class size= 94/10= 9.4

Number of class intervals = 10.

The minimum value = 804

The first-class interval is 804 – 814

Thus, the other class intervals are

814 – 824, 824 – 834, 834 – 844, 844 – 854, 854 – 864, 864 – 874,874 – 884, 884 – 894, 894 – 904.

So, we obtain the following frequency distribution table:

Frequency Distribution of Monthly Wages

Monthly Wages	Tally Marks	Frequency
804 – 814	𝍸	5
814 – 824	|	1
824 – 834	|||	3
834 – 844	𝍸 |||	8
844 – 854	||	2
854 – 864	||	2
864 – 874	||	2
874 – 884	|	1
884 – 894	𝍸	5
894 – 904	|	1
Total		**30**

Example 8: The following are weekly wages of 30 workers. Tabulate the data by grouping them in equal class intervals, one of them being 90 – 100 (100 not included):

96, 162, 108, 168, 138, 94, 128, 70, 88, 154, 134, 154, 88, 150, 86, 126, 150, 76, 62, 168, 158, 108, 102, 160, 76, 124, 150, 74, 94, 142.

Notes

Solution: The minimum and maximum weekly wages in the given raw data are 62 and 168. It is given that 90 – 100 is one of the class intervals and the class size in same. So, the classes of equal sizes are:

60 – 70, 70 – 80, 80 – 90, 90 – 100, 100 – 110, 110 – 120, 120 – 130, 130 – 140, 140 – 150, 150 – 160, 160 – 170.

Frequency Distribution of Weekly Wages

Weekly Wages (in Rs.)	Tally Marks	Frequency
60 – 70	|	1
70 – 80	||||	4
80 – 90	|||	3
90 – 100	|||	3
100 –110	|||	3
110 – 120	....	0
120 – 130	|||	3
130 – 140	||	2
140 – 150	|	1
150 – 160	$\not{\|\|\|\|}$ |	6
160 – 170	||||	4
Total		**30**

Note: The above frequency table is prepared by the exclusive method. The observation 150 is not included in the class 140 – 150 but is included in the next class 150 – 160. Similarly, 160 is included in the class 160 – 170.

Example 9: The class marks of a distribution are 61, 66, 71, 76, 81, 86, 91, 96, 101, 106. Determine the class size, class limits and true class limits.

Solution: Here the class marks are uniformly spaced, so the class size is the difference between any two consecutive class marks.

Therefore, class size = 66 – 61 = 5.

Let the lower limit of the first-class interval be a.

Then its upper limit = a + 5

Mid-value of the first interval = 61

Class mark = (Upper limit + Lower limit)/2

Notes

Then,

$[a + (a + 5)]/2 = 61$

$a + (a + 5) = 122$

$2a + 5 = 122$

$2a = 117$

$a = 58.5$

The first-class interval is 58.5 – 63.5.

and the other class intervals are 58.5 – 63.5, 63.5 – 68.5, 68.5 – 73.5, 73.5 – 78.5, 78.5 – 83.5, 83.5 – 88.5, 88.5 – 93.5, 93.5 – 98.5, 98.5 – 103.5 and 103.5 – 108.5.

Class Marks	Class Interval
61	58.5 – 63.5
66	63.5 – 68.5
71	68.5 – 73.5
76	73.5 – 78.5
81	78.5 – 83.5
86	83.5 – 88.5
91	88.5 – 93.5
96	93.5 – 98.5
101	98.5 – 103.5
106	103.5 – 108.5

Therefore, the classes are exclusive, so the true class limits are same as class limits. The lower-class limits are 58.5, 63.5, 68.5, 73.5, 78.5, 83.5, 88.5, 93.5, 98.5 and 103.5.

The upper-class limits are 63.5, 68.5, 73.5, 78.5, 83.5, 88.5, 93.5, 98.5, 103.5 and 108.5.

2.5 When the Mid Value of the Class and the Class Size are Given

Example 10: The following data gives the weights (in grams) of 50 oranges picked from a basket. Construct a grouped frequency distribution taking class-intervals of equal width 20 in such a way that the mid-value of the first-class interval is 10.

Notes

106, 107, 76, 82, 109, 107, 115, 93, 187, 95, 123, 125, 11, 92, 86, 70, 126, 68, 130, 129, 139, 119, 115, 128, 100, 18, 84, 99, 113, 204, 111, 141, 136, 123, 90, 115, 98, 110, 7, 90, 107, 81, 131, 75, 84, 104, 1

Solution: Here the size of each class = 20

Mid-value of the first class = 10

Let the lower limit of the first-class interval be a.

Then its upper limit = a + 20

Mid-value of the first interval = 10

Frequency Distribution = 21

(a + (a + 20))/2 = 10

a + a + 20 = 20

2a = 0

a = 0

Upper limit of the class = 0 + 20 = 20

Thus, First-class-interval is 0 – 20 and then other classes are:

20 – 40, 40 – 60, 60 – 80, 80 – 100, 100 – 120, 120 – 140, 140 – 160, 160 – 180, 180 – 200 and 200 – 220.

Frequency Distribution of Weights

Weights (in Grams)	Tally Marks	Frequency
0 – 20	|||	3
20 – 40	....	0
40 – 60	....	0
60 – 80	||||	4
80 – 100	𝍸 𝍸 ||||	14
100 – 120	𝍸 𝍸 𝍸 |	16
120 – 140	𝍸 𝍸	10
140 – 160	|	1
160 – 180	....	0
180 – 200	|	1
200 – 220	|	1
Total		**50**

Notes

Example 11: The weights in gms of 50 mangoes picked at random from a consignment are as follows:

141, 123, 92,85, 214, 91, 94, 128, 114, 120, 90, 117, 121, 151, 146, 133, 100, 88, 100, 125, 120, 108, 116, 109, 117, 94, 86, 196, 92, 110, 119, 138, 125, 117, 125, 129, 103, 197, 149, 139, 140, 78, 205, 133, 135, 121, 102, 96, 80, 136.

Form the grouped frequency table by dividing the variable range into intervals of equal width of 20 gms such that the mid-value of the first interval is 80 gms.

Solution: It is given that size of each class = 20.

Let the lower limit of the first-class interval be a. Then, its upper limit = (a + 20)

Mid-value of the first-class interval = 80

(a + (a + 20))/2= 80

2a + 20 = 160

2a = 140

a = 70

Thus, the first-class interval is 70 – 90 and the other classes are:

90 – 110, 110 – 130, 130 – 150, 150 – 170, 170 – 190, 190 – 210, 210 – 230

So, the frequency distribution table is as under:

Frequency Distribution of Weights of Mangoes

Weight (in gm)	Tally Marks	Frequency
70 – 90	𝍸	5
90–110	𝍸 𝍸 |||	13
110–130	𝍸 𝍸 𝍸 ||	17
130 – 150	𝍸 𝍸	10
150 – 170	|	1
170 – 190	....	0
190 – 210	|||	3
210 – 230	|	1
Total		**50**

2.6 Cumulative Frequency

Definition: The cumulative frequency corresponding to a class is the sum of all the frequencies up to and including that class. Let us consider marks obtained by 50 students in M.Sc. in a test.

1 student secured zero marks.

3 students, each secured 5 marks.

9 students, each secured 10 marks and so on.

How many students secured 10 marks or less marks?

Then we have to add all the frequencies corresponding to 0 marks, 5 marks and 10

i.e., (1 + 3 + 9) students = 13 students.

It means 13 students secured 10 marks or less than 10 marks. 13 is termed as cumulative frequency for marks 10.

i.e., 1 + 3 + 9 + 11 = 24

It means 24 students secured 14 marks or less than 14 marks.

24 is termed as cumulative frequency for marks 14.

2.7 Types of Cumulative Frequencies

There are two types of cumulative frequencies. (i) Less than series (ii) More than series.

For less than cumulative frequencies we add up the frequencies from above.

For more than cumulative frequencies we add up the frequencies from below.

Example 12: Construct cumulative frequency distribution (less than series and more than series) from the following data:

Marks Obtained	0 – 20	20 – 40	40 – 60	60 – 80	80 – 100
No. of Students	2	7	11	18	12

Solution:

(i) Less than cumulative frequency table

Cumulative Frequency Distribution Table

Marks Obtained	No. of Students (Cumulative Frequency)
Less than 20	2
Less than 40	9 = 2 + 7

Notes

Less than 60	20 = 2 + 7 + 11
Less than 80	38 = 2 + 7 + 11 + 18
Less than 100	50 = 2 + 7 + 11 + 18 + 12

(ii) More than cumulative frequency table

Cumulative Frequency Distribution Table

Marks Obtained	No. of Students (Cumulative Frequency)
More than 0	50 = (12 + 18 + 11 + 7 + 2)
More than 19	48 = (12 + 18 + 11 + 7)
More than 39	41 = (12 + 18 + 11)
More than 59	30 = (12 + 18)
More than 79	12

Example 13: Following are the ages in years of 360 patients getting medical treatment in a hospital:

Age (in years)	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of Patients	90	50	60	80	50	30

Construct the cumulative frequency table (less than series and more than series) for the above data.

Solution:

Cumulative Frequency Distribution Table (Less Than Type)

Age (in years)	No. of Patients
less than 20	90
less than 30	140 = (90 + 50)
less than 40	200 = (90 + 50 + 60)
less than 50	280 = (90 + 50 + 60 + 80)
less than 60	330 = (90 + 50 + 60 + 80 + 50)
less than 70	360 = (90 + 50 + 60 + 80 + 50 + 30)

Cumulative Frequency Distribution Table (More Than Type)

Age (in Years)	No. of Patients
More than 9	360 = (30 + 50 + 80 + 60 + 50 +90)
More than 19	270 = (30 + 50 + 80 + 60 + 50)
More than 29	220 = (30 + 50 + 80 + 60)

Notes

More than 39	160 = (30 + 50 + 80)
More than 49	80 = (30 + 50)
More than 59	30

Example 14: A cumulative frequency distribution is given below. Convert this into a frequency distribution table.

Marks	Less than 30	Less than 40	Less than 50	Less than 60	Less than 70	Less than 80	Less than 90	Less than 100
No. of Students	5	8	20	24	31	40	46	48

Solution: To form a frequency distribution from cumulative frequency distribution, we subtract the cumulative frequency of preceding class from the cumulative frequency of each class. The result gives us the frequency for that class. A required frequency distribution table is given below:

Marks	**No. of Students**
0 – 30	5
30 – 40	3 = 8 – 5
40 – 50	12 = 20 – 8
50 – 60	4 = 24 – 20
60 – 70	7 = 31 – 24
70 – 80	9 = 40 – 31
80 – 90	6 = 46 – 40
90 – 100	2 = 48 – 46

Example 15: Make a frequency table from the following:

Age	More than 100	More than 90	More than 80	More than 70	More than 60	More than 50	More than 40	More than 30	More than 20	More than 10	More than 0
No. of Persons	0	7	18	26	41	63	75	98	117	139	150

Solution:

Age	**No. of Persons**
0 – 10	11 = 150 – 139
10 – 20	22 = 139 – 117
20 – 30	19 = 117 – 98
30 – 40	23 = 98 – 75
40 – 50	12 = 75 – 63

Notes

50 – 60	22 = 63 – 41
60 – 70	15 = 41 – 26
70 – 80	8 = 26 – 18
80 – 90	11 = 18 – 7
90 – 100	7 = 7 – 0

Example 16: Find the unknown entries (a, b, c, d, e, f, g) in the following frequency distribution table of heights of 50 students in a class:

Class (Heights in cm)	Frequency	Cumulative Frequency
150 – 155	12	a
155 – 160	b	25
160 – 165	10	c
165 – 170	d	43
170 – 175	e	48
175 – 180	2	f
	g	**50**

Solution: Cumulative Frequency Distribution Table (More than type)

x Class (Height in cm)	Frequency	Cumulative Frequency
less than 155	12	12 = a
less than 160	b	12 + b = 25
less than 165	10	12 + b + 10 = c
less than 170	d	12 + b + 10 + d = 43
less than 175	e	12 + b + 10 + d + e = 48
less than 180	2	12 + b + 10 + d + e + 2 = f
	g	**50**

a = 12

12 + b = 25

b = 25 – 12 = 13

12 + b + 10 = c

12 + 13 + 10 = c

c = 35

12 + b + 10 +d = 43

12 + 13 + 10 + d = 43,	or	d = 43 – 12 – 13 – 10 = 8
12 + b + 10 +d + e = 48	or	12 + 13 + 10 + 8 + e = 48
43 + e = 48	or	e = 5

12 + b + 10 + d + e + 2 = f	or	12 + 13 + 10 + 8 + 5 + 2 = f
50 = f		

g = 50

Hence, a = 12, b= 13, c = 35, d = 8, e =5, f = 50, g = 50

2.8 Miscellaneous Questions

Example 17: A sample consists of 34 observations recorded correct to the nearest integer, ranging in value from 201 to 337. If it is decided to use seven classes of width 20 integers and to begin in the first class at 199.5, find the class limits and class marks of the seven classes.

Solution: The class limits will be the range of values that each class covers, and the class marks will be the midpoint of each class interval.

Class Interval (h): The class interval is given as 20.

Starting Point: The first class starts at 199.5.

Number of Classes (k): The number of classes is given as 7.

To find the class limits and class marks, we will follow these steps:

- Lower Limit of the First Class (L_1): 199.5
- Upper Limit of the First Class (U_1): $L_1 + h = 199.5 + 20 = 219.5$
- Lower Limit of the Second Class (L_2): 219.5
- Upper Limit of the First Class (U_2): $L_2 + h = 219.5 + 20 = 239.5$

For the subsequent classes: Lower Limit of the $(i + 1)^{th}$ Class (L_{i+1}): U_i and Upper Limit of the $(i+1)^{th}$ Class (U_{i+1}): $L_{i+1}+h$

The class mark (M) for each class is calculated by adding lower limits and upper limits and dividing it by 2. For the first Class, Class Mark will be (199.5 + 219.5)/2 = 209.5

The following table gives the class limits and class marks of the seven classes.

Class	Lower Limit	Upper Limit	Class Mark
1	199.5	219.5	209.5
2	219.5	239.5	229.5

Notes

3	239.5	259.5	249.5
4	259.5	279.5	269.5
5	279.5	299.5	289.5
6	299.5	319.5	309.5
7	319.5	339.5	329.5

Example 18: In a trip organized by a college there were 80 persons, each of whom paid Rs. 15.50 on an average. There were 60 students each of whom paid Rs. 16. Members of the teaching staff charged at a higher rate. The number of servants were 6 (all males) and they were not charged anything. The number of ladies was 20% of the total of which one was lady staff member. Tabulate the above information.

Solution:

Table Showing the Type of Participants, Sex, and Contribution Made

	Sex			Contribution per Member	Total Contribution
Type of Participants	**Males**	**Females**	**Total**	**(Rs.)**	**(Rs.)**
Students	45	15	60	16.00	960
Teaching Staff	13	1	14	20.00	280
Servants	6	-----	6	-----	-----
Total	**64**	**16**	80		**1240**

Notes:

1. Total Contribution = Average contribution × No. of persons joined the trip

$$= 15.5 \times 80 = 1240$$

2. Contribution of the staff per head has been obtained by deducting the contribution of students from the total and dividing the difference by the number of teaching staff i.e.

$$\frac{1240-(60\times16)}{14}=\frac{1240-960}{14}=\frac{280}{14}=Rs.20$$

Example 19: A survey of 370 students from Commerce Faculty and 130 students from Science Faculty revealed that 180 students were studying for only C.A. Examination, 140 for only Costing Examination and 80 for both C.A. and Costing Examinations. The rest had offered part-time

Notes

Management Course. Of those studying for Costing only, 13 were girls and 90 boys belong to Commerce Faculty. Out of 80 studying for both C.A. and Costing 72 were from Commerce Faculty amongst which 70 were boys. Amongst those who offered part-time Management Course, 50 boys were from Science Faculty and 30 boys and 10 girls from Commerce Faculty. In all there were 110 boys in Science Faculty. Present the above information in tabular form. Find the number of students studying for part-time Management Course.

Solution:

Distribution of Students According to Professional Course

	Faculty								
Courses	**Commerce Boys**	**Commerce Girls**	**Commerce Total**	**Science Boys**	**Science Girls**	**Science Total**	**Total Boys**	**Total Girls**	**Grand Total**
Part-time Management	30	10	40	50	10	60	80	20	**100**
Only C.A.	**150**	**8**	**158**	**16**	**6**	**22**	**166**	**14**	180
Only Costing	90	10	103	37	3	40	127	13	140
C.A. and Costing	70	**2**	72	7	1	8	77	3	80
Total	**340**	**30**	370	**110**	**20**	130	450	50	**500**

Note on Calculations:

- Total number of students = 370 (Commerce) + 130 (Science) = 500
- Students studying in part-time management courses = 500 – (180 + 140 + 80) = 500 – 400 = 100

Example 20: Prepare a frequency table for the following data with width of each class-interval as 10. Use exclusive method of classification.

57	44	80	75	0	18	45	14	4	64
72	51	69	34	22	83	70	20	57	28
96	56	50	47	10	43	61	66	80	46
22	10	84	50	47	73	42	33	48	65
10	34	66	53	75	90	58	46	38	69

Solution:

Preparation of Frequency Distribution

Variable	Tally Bars	Frequency
0–10	||	2
10–20	𝍸	5

Notes

20–30	||||	4
30–40	𝍸	5
40–50	𝍸 |||	8
50–60	𝍸 |||	8
60–70	𝍸 ||	7
70–80	𝍸	5
80–90	||||	4
90–100	||	2
		Total = 50

Example 21: Prepare a continuous frequency distribution from the following observation:

75	42	70	37	62	70	50
60	45	81	56	31	45	25
31	62	78	80	78	56	55
75	58	72	32	50	26	70
15	55	40	68	35	60	60
42	81	43	69	65	62	58
42	80	40	45	75	45	62

Solution: Since the lowest value is 15 and largest is 81 and suppose we want to take 10 class intervals.

Continuous Frequency Distribution

Class Intervals	Tallies	Frequency
15-25	|	1
25-35	𝍸	5
35-45	𝍸 |||	8
45-55	𝍸 |	6
55-65	𝍸 𝍸 |||	13
65-75	𝍸 ||	7
75-85	𝍸 ||||	9
		Total = 49

Example 22: Count the number of letters in each word of the para given below (ignore punctuation marks). Prepare the discrete frequency distribution.

Notes

"There is no problem which I and God cannot solve together."

"There is no situation which I and God cannot handle together."

"There is no burden which I and God cannot bear together."

Solution:

Step 1:

1. **First Sentence:** "There is no problem which I and God cannot solve together."
   - **Words:** There, is, no, problem, which, I, and, God, cannot, solve, together
   - **Letter Counts:** 5, 2, 2, 7, 5, 1, 3, 3, 6, 5, 8
2. **Second Sentence:** "There is no situation which I and God cannot handle together."
   - **Words:** There, is, no, situation, which, I, and, God, cannot, handle, together
   - **Letter Counts:** 5, 2, 2, 9, 5, 1, 3, 3, 6, 6, 8
3. **Third Sentence:** "There is no burden which I and God cannot bear together."
   - **Words:** There, is, no, burden, which, I, and, God, cannot, bear, together
   - **Letter Counts:** 5, 2, 2, 6, 5, 1, 3, 3, 6, 4, 8

Step 2: Create Discrete Frequency Distribution

We will now count the frequency of each letter count across all three sentences.

- 1 letter: I (3 times)
- 2 letters: is, no (6 times)
- 3 letters: and, God (6 times)
- 4 letters: bear (1 time)
- 5 letters: There, which, solve (7 times)
- 6 letters: cannot, burden, handle (5 times)
- 7 letters: problem (1 time)
- 8 letters: together (3 times)
- 9 letters: situation (1 time)

Notes

Discrete Frequency Distribution

Number of Letters	Tally	Frequency
1	|||	3
2	𝍸 |	6
3	𝍸 |	6
4	|	1
5	𝍸 ||	7
6	𝍸	5
7	|	1
8	|||	3
9	|	1
		Total = 3

Example 23: Following are the numbers of items of similar type produced in a factory during the last 50 days.

21	22	17	23	27	15	16	22	15	23
24	25	36	19	14	21	24	25	14	18
20	31	22	19	18	20	21	20	36	18
21	20	31	22	19	18	20	20	24	35
25	26	19	32	22	26	25	26	27	22

Arrange these observations in to a frequency distribution with both inclusive and exclusive class intervals choosing a suitable number of classes.

Solution:

Range Calculation: Minimum value: 14 and Maximum value: 36. Therefore, Range: 36 – 14 = 22

Since the number of observations are 50, it seems reasonable to choose 6 (2^6 > 50) or less classes. The class interval is given by

$$h = \frac{Range}{Number\,of\,classes} = \frac{22}{6} = 3.66 \; or \; 4$$

Performing the actual tally and counting the number of observations in each class, we get the following frequency distribution with inclusive class intervals as shown in Table 2.1.

Notes

Table 2.1: Frequency Distribution with Inclusive Class Intervals

Class Intervals	Tally	Frequency (Number of Items Produced)
14 - 17	𝍸 |	6
18 - 21	𝍸 𝍸 𝍸 |||	18
22 - 25	𝍸 𝍸 𝍸	15
26 - 29	𝍸	5
30 - 33	|||	3
34 - 37	|||	3
Total		**50**

Converting the class intervals shown in Table 2.1 in to exclusive class intervals is shown in Table 2.2.

Table 2.2: Frequency Distribution with Exclusive Class Intervals

Class Intervals	Mid-Value of Class Intervals	Frequency (Number of Items Produced)
13.5 – 17.5	15.5	6
17.5 – 21.5	19.5	18
21.5 – 25.5	23.5	15
25.5 – 29.5	27.5	5
29.5 – 33.5	31.5	3
33.5 – 37.5	34.5	3

Example 24: Prepare a statistical table from the following weekly wages of 100 workers (in Rs.) of Factory A:

88	23	27	28	86	96	94	93	86	99
82	24	24	55	88	99	55	86	82	36
96	39	26	54	87	100	56	84	83	86
102	48	27	26	29	100	59	83	84	48
104	46	30	29	40	101	60	89	46	49
106	33	36	30	40	103	70	90	49	50
104	36	37	40	46	108	72	24	50	60
24	39	49	46	66	107	76	96	46	67
26	78	50	44	43	46	49	99	96	68
29	67	56	99	93	48	80	102	32	51

Solution: The lowest value is 23 and the highest 105. The difference in the highest and the lowest value is 83. If we take a class interval of 10,

Notes

nine classes would be formed. The first class would be taken as 20–30 instead of 23–33 as per the principles of classification.

Wages (Rs.)	Tally	Frequency
20-30	𝍸 𝍸 |||	13
30-40	𝍸 𝍸 |	11
40-50	𝍸 𝍸 𝍸 |||	18
50-60	𝍸 𝍸	10
60-70	𝍸 |	6
70-80	𝍸	5
80-90	𝍸 𝍸 ||||	14
90-100	𝍸 𝍸 ||	12
100-110	𝍸 𝍸 |	11
		Total = 100

Example 25: Present the following data of the percentage marks of 60 students in the form of a frequency table with 10 classes of equal widths, one class being 50-59.

41	17	83	63	54	92	60	58	70	06	67	82
33	44	57	49	34	73	54	63	36	52	32	75
60	33	09	72	28	30	42	93	43	80	03	32
57	67	24	64	63	11	35	82	10	23	00	41
60	32	72	53	92	88	62	55	60	33	40	57

Solution:

Formulation of Frequency Distribution

Marks	Tally Bars	Frequency
0-9	||||	4
13-19	|||	3
20-29	|||	3
30-39	𝍸 𝍸	10
40-49	𝍸 |||	7
50-59	𝍸 ||||	9
60-69	𝍸 𝍸 |	11
70-79	𝍸	5
80-89	𝍸	5
90-99	|||	3
		Total = 60

Notes

2.9 Exercise

1. Explain the cumulative frequency distribution.

2. What is the difference between a frequency distribution and a cumulative frequency distribution.

3. The following is the distribution of weights (in kg) of 40 persons:

Weight (in kg)	No. of Persons
40 – 45	4
45 – 50	4
50 – 55	13
55 – 60	5
60 – 65	6
65 – 70	5
70 – 75	2
75 – 80	1
Total	**40**

(i) Determine the class marks of the class 40 – 45, 45 – 50, etc.

(ii) Construct the cumulative frequency distribution table. (Less than type and more than type)

4. The following is the distribution of marks of 180 primary school students of Allahabad:

Frequency Distribution Table

Marks	No. of Students
Less than 20	10
20 – 25	33
25 – 30	52
30 – 35	47
35 – 40	28
40 – 45	6
45 – 50	4
Total	**180**

Construct a cumulative frequency distribution table. (Less than type)

5. In a study of diabetic patients, the following data are obtained.

Notes

Frequency Distribution Table

Age (in years)	No. of Patients
10 – 20	6
20 – 30	12
30 – 40	28
40 – 50	18
50 – 60	10
60 – 70	4

Construct a cumulative frequency table, less than type and more than type for the above data.

6. Make a frequency table for the following cumulative frequency distribution table:

Marks	No. of Students
Less than 10	2
Less than 20	8
Less than 30	12
Less than 40	19
Less than 50	31
Less than 60	42
Less than 70	60
Less than 80	75
Less than 90	88
Less than 100	90

7. The following cumulative frequency distribution table shows the ages of people living in a locality.

Cumulative Frequency Distribution Table

Age (in years)	No. of Persons
Above 108	0
Above 96	7
Above 84	9
Above 72	11
Above 60	26
Above 48	164
Above 36	433
Above 24	815
Above 12	1032
Above 0	1130

Construct a frequency table.

Notes

8. In a study of diabetic patients, the following data are obtained.

Frequency Distribution Table

Age (in years)	No. of Patients
10 – 20	3
20 – 30	18
30 – 40	25
40 – 50	20
50 – 60	15
60 – 70	9

Construct a cumulative frequency table, less than type and more than type for the above data.

9. Make a frequency table for the following cumulative frequency distribution table:

Marks	No. of Students
Less than 10	2
Less than 20	8
Less than 30	12
Less than 40	19
Less than 50	31
Less than 60	42
Less than 70	60
Less than 80	75
Less than 90	88
Less than 100	90

10. The following cumulative frequency distribution table shows the ages of people living in a locality.

Cumulative Frequency Distribution Table

Age (in years)	No. of Persons
Above 108	0
Above 96	7
Above 84	9
Above 72	11
Above 60	26
Above 48	164

Notes

Above 36	433
Above 24	815
Above 12	1032
Above 0	1130

Construct a frequency table.

11. Present the following information in a suitable tabular form:

In 1965 out of a total of 1,750 workers of a factory, 1,200 were members of a trade union. The number of women employed was 200 of which 175 did not belong to a trade union. In 1970 the number of union workers increased to 1,580, of which 1,290 were men. On the other hand, the number of non-union workers fell down to 208, of which 180 were men. In 1975, there were 1,800 employees who belonged to a trade union and 50 who did not belong to a trade union. Of all the employees in 1975, 300 were women of whom only 8 did not belong to a trade union.

12. In a sample study about the coffee habits in two towns, the following data were observed:

Town A	60% people were males, 40% were coffee drinkers, and 26% were male coffee drinkers.
Town B	55% people were males, 30% were coffee drinkers, and 20% were male coffee drinkers.

Tabulate the above observations.

13. From the following observations, prepare a frequency distribution table in ascending order starting with 4000–4100 (exclusive method):

Income (in Rs.)

4252	4083	4125	4262	4103	4324	4366	4303
4201	4307	4363	4381	4256	4118	4125	4252
4472	4372	4454	4500	4425	4357	4374	4321
4942	4556	4404	4489	4325	4547		

Notes

14. In the annual report of a mobile oil company, it is indicated that the company drilled a total of 882 wells in 2008 and 487 in 2009. Two types of drilling operations were conducted: Wildcat and developmental. In 2008, a total of 40 wildcat wells and 842 developmental wells were drilled. The comparable figures for 2009 were 46 and 441. There were 3 possible outcomes when a well was drilled: oil, gas, or dry hole. Of the wildcat wells drilled in 2008, 6 resulted in oil, 4 in gas, 30 in dry holes. The comparable figures for 2009 were 6, 4, and 36. Of the developmental wells drilled in 2008, 660 resulted in oil, 77 in gas, and 105 in dry holes; the comparable figures for 2009 were 333, 44, and 64.

Present the information in the above paragraph in a formal table giving an appropriate title.

15. The weight in grams of 50 apples, picked from a box are as follows:

110	103	89	75	98	121	110	108	93	128
185	123	113	92	86	70	126	78	139	120
129	119	105	120	100	116	85	99	114	189
205	111	141	136	123	90	115	128	160	78
90	107	81	137	75	84	104	109	87	115

Construct a frequency table with a class interval of 15 grams.

16. A portfolio contains 51 stocks whose prices are given below:

67	34	36	48	49	31	61	34
43	45	38	32	27	61	29	47
36	50	46	30	40	32	30	33
45	49	48	41	53	36	37	47
47	30	50	28	35	35	38	36
46	43	34	62	69	50	28	44
43	60	39					

Summarize these stock prices in the form of a frequency distribution.

17. Construct a frequency distribution of the data given below, where class interval is 4 and the mid-value of one of the classes is zero.

-8	-7	10	12	6	4	3	0	7
-4	-3	-2	2	3	4	7	5	6
10	12	9	13	11	-10	-7	1	0
5	3	2	6	10	-6	-4		

Notes

18. Classify the following data by taking class intervals such that their mid-values are 17, 22, 27, 32, and so on:

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

19. Following are the number of two wheelers sold by a dealer during eight weeks of six working days each.

13	19	22	14	13	16	19	21
23	11	27	25	17	17	13	20
23	17	26	20	24	15	20	21
23	17	29	17	19	14	20	20
10	22	18	25	16	23	19	23
21	17	18	24	21	20	19	26

(i) Group these figures into a table having the classes 10–12, 13–15, 16–18, ..., and 28–30.

(ii) Convert the distribution of part (i) into a corresponding frequency distribution and also a cumulative frequency distribution.

20. Of the 1,125 students studying in a co-ed college during a year, 720 were Hindus, 628 were boys, and 440 were science students; the number of Hindu boys was 392, that of boys studying science 205 and that of Hindus students studying science 262; the number of science students among the Hindu boys was 148. Enter these frequencies in a three-way table with the rows representing the Faculty (Science and Arts), and the columns representing Religion (Hindus and Non-Hindus) and Gender (Boys and Girls and complete the table by obtaining the frequencies of the remaining) cells.

LESSON

3

Histogram, Frequency Polygons, Frequency Curves and Ogives

STRUCTURE

3.1 Histogram

Histogram is a graphical representation of a grouped frequency distribution with continuous class. It consists of a set of rectangles having their heights proportional to their class frequencies, for equal class intervals. There is no gap between two successive rectangles. The rectangles are constructed with base as the class size and their heights representing the frequencies.

Drawing of Histogram for Continuous Grouped Frequency Distribution

1. Along the x-axis class intervals are marked.
2. The corresponding frequencies are marked on the y-axis.
3. The rectangles are constructed with class intervals as bases and the corresponding frequencies as heights.

Note:

If the class intervals are not continuous then they are to be converted into continuous distribution by the following method:

Notes

Let h = (Lower limit of II class interval – Upper limit of I class interval)

Then subtract h/2 from the lower limits of each class and add h/2 to the upper limits of each class.

If the mid-points of class intervals are given, then compute the difference between the second and first mid-point.

Let this difference = h, find h/2

Then subtract from each mid-point to get the lower limit of each class and add h/2 to each mid-point to get upper limit of each class.

Suppose first class interval start from 20 and not from zero. We show it on the graph by making a ''kink'' or a break on the axis.

3.2 Bar Graph

Bar graphs are the pictorial representation of data (generally grouped), in the form of vertical or horizontal rectangular bars, where the length of bars is proportional to the measure of data. They are also known as bar charts. Bar graphs are one of the means of data handling in statistics.

Steps to Draw Bar Graph

In order to visually represent the data using the bar graph, we need to follow the steps given below:

Step 1: First, decide the title of the bar graph.

Step 2: Draw the horizontal axis and vertical axis.

Step 3: Now, label the horizontal axis.

Step 4: Write the names on the horizontal axis.

Step 5: Now, label the vertical axis.

Step 6: Finalise the scale range for the given data.

Step 7: Finally, draw the bar graph that should represent each category with their respective numbers.

Difference Between Bar Graph and Histogram

- In histogram there is no gap in between consecutive rectangles as in bar graph.
- The width of the bar is significant in histogram. In bar graph, width is not important at all.

Notes

- In histogram the areas of the rectangles are proportional to the frequency, however if the class size of the frequencies are equal then each of the rectangle are proportional to the frequencies.

Example 1: A survey conducted by an organisation in respect of illness and death among the women between the ages 15 – 44 (in years) world-wide, found the following figures (in%):

S. No.	Causes	Female Fatality Rate (%)
1.	Reproductive health conditions	31.8
2.	Neuropsychiatric conditions	25.4
3.	Injuries	12.4
4.	Cardiovascular conditions	4.3
5.	Respiratory conditions	4.1
6.	Other causes	22.0

(i) Represent the information given above graphically.

(ii) Which condition is the major cause of women's ill health and death worldwide?

(iii) Try to find out, any two factors which play a major role in the cause in (ii)above being the major cause.

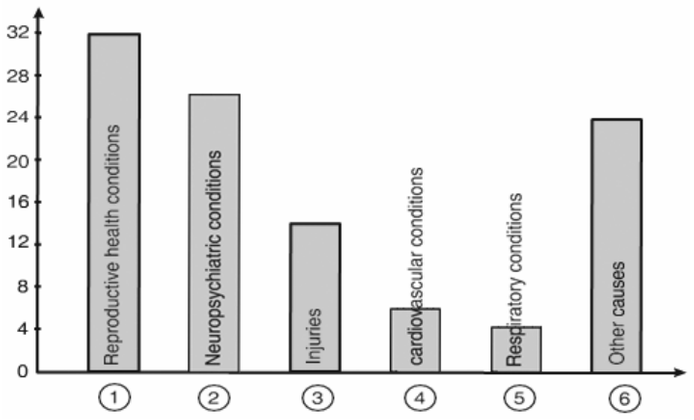
Solution:

(i) We draw the bar graph of the given data in the following steps:

The variable here is the causes of death and value of the variable is female fatality rate in percentage.

1. We represent the causes of death on horizontal axis.
2. The width of the bar is not important at all but we take equal width for all the bars and maintain equal gaps between.
3. We represent the female fatality rate on the vertical axis. The maximum female fatality rate is 31.8%. We take scale as1 cm = 4%.
4. We draw first rectangular bar with width 1 cm and height 7.45 cm.
5. We draw second rectangular bar with a gap of 1 cm in between 1st and 2nd bar and the height of the bar is 6.35 cm.
6. Similarly, other bars are also drawn as we have done in step 3 and

Notes



(ii) Reproductive health condition is the major cause of women's ill health and death worldwide.

(iii) Three factors which play a major role is:

1. Number of maternity hospitals is less, where the pregnant women get advice about delivery, etc. from qualified doctors.
2. Unwanted female child.
3. Abortion by unqualified doctors.

Example 2: The following data on the number of girls to the nearest ten per thousand boys in different sections of the society is given below:

Section	Number of Girls Per Thousand Boys
Scheduled Caste (SC)	940
Scheduled Tribe (ST)	970
Non SC/ST	920
Backward districts	950
Non-backward districts	920
Rural	930
Urban	910

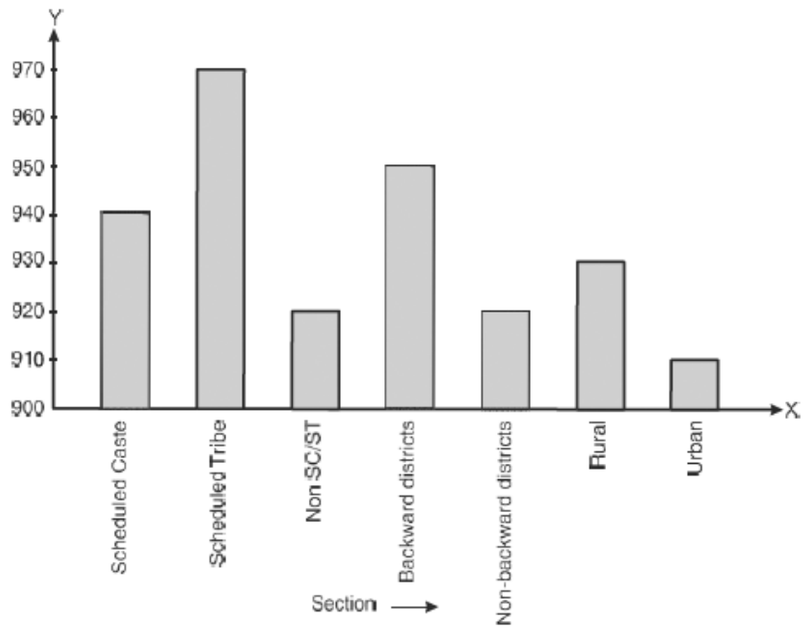
(i) Represent the information above by a bar graph.

(ii) Write two conclusions you can arrive at from the graph, with justification.

Solution:

(i) The required graph is under:

Notes



(ii) In the graph drawn, different sections of the society are denoted along the horizontal axis and the number of girls to the nearest ten per thousand boys are denoted along the vertical axis, their intersection represents 900.

Scale: 1 cm = 10 girls.

From the graph, we find that the number of girls to the nearest ten per thousand boys are maximum in scheduled tribes, whereas they are minimum in urban.

Example 3: The daily earnings of 50 workers are given below:

Daily Earnings (in Rs.)	125–134	135–144	145–154	155–164	165–174	175–184
No. of Workers	2	7	10	15	10	6

Draw a histogram.

Solution: Since the given data is not continuous, so we have to convert them into continuous frequency distribution.

Let h = lower limits of II class – upper limit of I class

h = 135 – 134 = 1

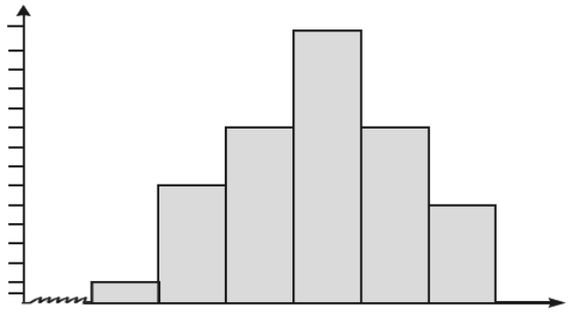
h/2 = 0.5

Notes

We subtract 0.5 from the lower limits of each class and add 0.5 to the upper limits of each class. We get the following continuous frequency distribution:

Daily Earnings (in thousands)		No. of Workers
124.5	134.5	2
134.5	144.5	7
144.5	154.5	10
154.5	164.5	15
164.5	174.5	10
174.5	184.5	6

The graph is now plotted by taking Daily earnings on horizontal axis and No. of workers on vertical axis.



Example 4: Construct a histogram from the following distribution of total marks obtain by 65 students of Graduation in half yearly examination.

Marks (Mid-points)	75	80	85	90	95	100
No. of Students	7	11	25	13	6	3

Solution: The difference between the second and first mid-point

= 5 = h (say)

= 80 – 75

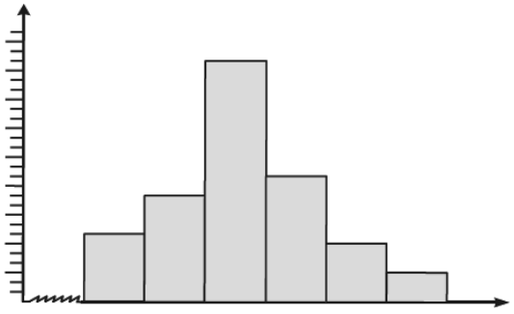
∴ h/2 = 5/2 = 2.5

Now subtract 2.5 from each mid-point to get the lower limits and add 2.5 to each mid-point to get the upper limits.

∴ We get the following frequency distribution:

Notes

Marks	72.5–77.5	77.5–82.5	82.5–87.5	87.5–92.5	92.5–97.5	97.5–102.5
No. of Students	7	11	25	13	6	3



3.3 Frequency Polygon

It is a line graph of class frequency plotted against class mark. It can be obtained by two methods: (i) By using Histogram (ii) Without using Histogram:

Steps of Drawing Frequency Polygon (By Using Histogram)

It can be obtained by connecting mid-points of the top of the rectangles of a histogram.

Step I: Draw the histogram from the given data.

Step II: Obtain the mid-points of the upper horizontal sides of each rectangle.

Step III: Join these mid-points of the adjacent rectangles by dotted lines.

Step IV: Obtain the mid-points of two class intervals of zero frequency, one adjacent to the first on its left and another adjacent to the last, on its right.

Step V: Complete the polygon by joining the mid-points of first and last class intervals to the mid-point of the imagined class intervals adjacent to them.

Example 5: For the following data, draw a histogram and a frequency polygon.

Age (in years)	0 – 6	6 – 12	12 – 18	18 – 24	24 – 30	30 – 36	36 – 42
No. of Persons	6	11	25	35	18	12	6

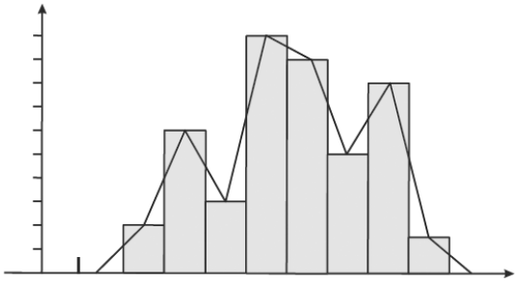
Notes

Solution: First, we draw the histogram of the given data. Then we will find the mid- points of the top of rectangles. Join these mid-points by dotted straight lines. Complete the polygon by joining the mid-points of first and last class intervals to the mid-points of imagined class intervals, with zero frequency, adjacent to them.

Example 6: For the following data, draw a histogram and a frequency polygon.

Marks	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100
No. of Students	5	12	6	20	18	10	16	3

Solution: First, we draw a histogram from the given data, then join the mid-points of the top of the rectangles by dotted straight lines. Complete the polygon by joining the mid-points of first and last class intervals to the mid-point of imagined class intervals adjacent to them.



Steps of Drawing Frequency Polygon, Without Using Histograms

Step I: Calculate the class marks (mid points of class intervals) $x_1, x_2, \ldots, x_n$ of the given class intervals. Class mark = (Upper limit + Lower limit)/2

Step II: Mark $x_1, x_2, \ldots, x_n$ along x-axis.

Step III: Mark the frequencies $f_1, f_2, \ldots, f_n$ along y-axis.

Step IV: Plot the points $(x_1, f_1), (x_2, f_2), \ldots, (x_n, f_n)$.

Notes

Step V: Join the points (x_1, f_1), (x_2, f_2), …, (x_n, f_n) by the line segments.

Step VI: Take two class intervals of zero frequency, one at the beginning and other at the end. Obtain their mid-points.

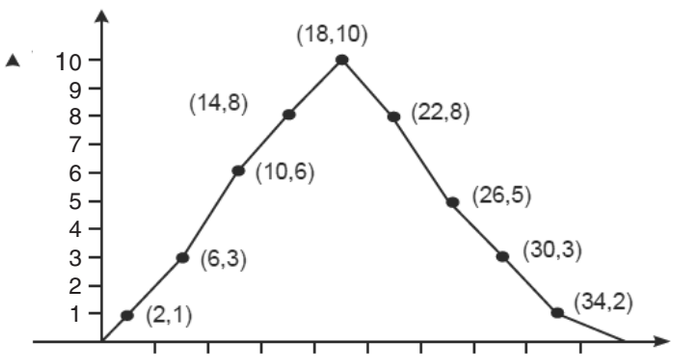
Step VII: Complete the frequency polygon by joining the mid-points of the first and the last intervals to the mid-point of the imagined classes adjacent to them.

Example 7: Construct a frequency polygon for the following data:

Age (in years)	0–4	4–8	8–12	12–16	16–20	20–24	24–28	28–32	32–36
No. of Persons	1	3	6	8	10	8	5	3	2

Solution:

Age	Class-Marks	No. of Persons
0 – 4	2	1
4 – 8	6	3
8 – 12	10	6
12 – 16	14	8
16 – 20	18	10
20 – 24	22	8
24 – 28	26	5
28 – 32	30	3
32 – 36	34	2



Example 8: The following table gives the distribution of I.Q.'s (Intelligence quotients) of 60 pupils of Hindu College.

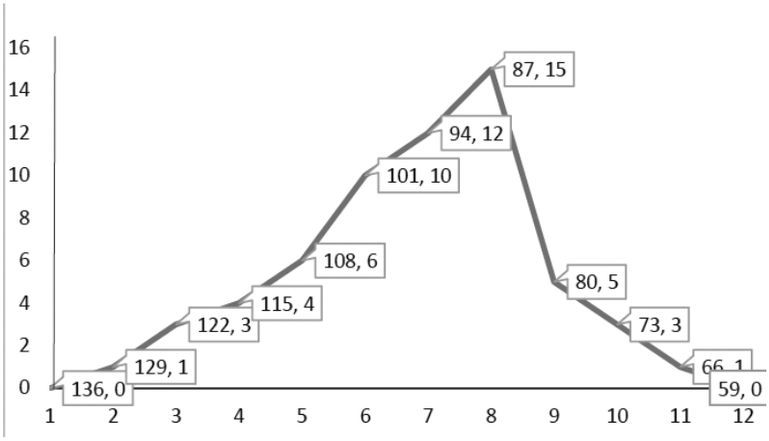
Notes

I.Q.	125.5–132.5	118.5–125.5	111.5–118.5	104.5–111.5	97.5–104.5	90.5–97.5	83.5–90.5	76.5–83.5	69.5–76.5	62.5–69.5
No. of Pupils	1	3	6	8	10	8	5	3	2	1

Construct a frequency polygon for above data.

Solution:

I.Q.	Class-Marks	No. of Pupils
125.5 – 132.5	129	1
118.5 – 125.5	122	3
111.5 – 118.5	115	4
104.5 – 111.5	108	6
97.5 – 104.5	101	10
90.5 – 97.5	94	12
83.5 – 90.5	87	15
76.5 – 83.5	80	5
69.5 – 76.5	73	3
62.5 – 69.5	66	1



Example 9: The runs scored by two teams A and B on the first 60 balls in a cricket match are given below:

Number of Balls	Team A	Team B
1 – 6	2	5
7 – 12	1	6
13 – 18	8	2

Notes

Number of Balls	Team A	Team B
19 – 24	9	10
25 – 30	4	5
31 – 36	5	6
37 – 42	6	3
43 – 48	10	4
49 – 54	6	8
55 – 60	2	10

Represent the data of both the teams on the same graph by frequency polygons.

Solution: First, we obtain the class marks as given in the following table:

Number of Balls	Class Marks	Number of Runs	
		Team A	Team B
1 – 6	3.5	2	5
7 – 12	9.5	1	6
13 – 18	15.5	8	2
19 – 24	21.5	9	10
25 – 30	27.5	4	5
31 – 36	33.5	5	6
37 – 42	39.5	6	3
43 – 48	45.5	10	4
49 – 54	51.5	6	8
55 – 60	57.5	2	10

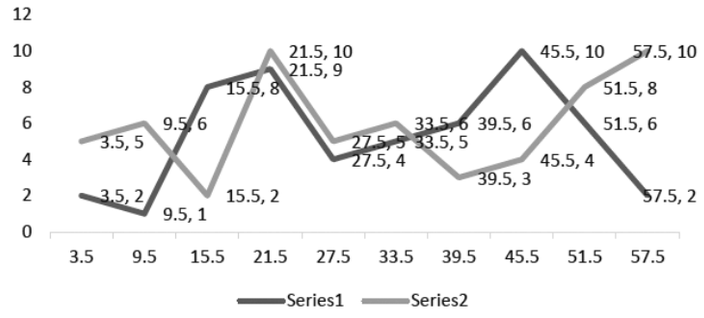
We represent class marks along x-axis on a suitable scale and frequencies (number of runs) along y-axis on a suitable scale.

To obtain the frequency polygon of Team A, we plot the points (3.5, 2), (9.5, 1), (15.5, 8), (21.5, 9), (27.5, 4), (33.5, 5), (39.5, 6), (45.5, 10), (51.5, 6) and (57.5, 2) and join these points by the line segments.

To obtain the frequency polygon of Team B, we plot the points (3.5, 5), (9.5, 6), (15.5, 2), (21.5, 10), (27.5, 5), (33.5, 6), (39.5, 3), (45.5, 4), (51.5, 8) and (57.5, 10).

The two frequency polygons are as shown below:

Notes



3.4 Pie Diagrams

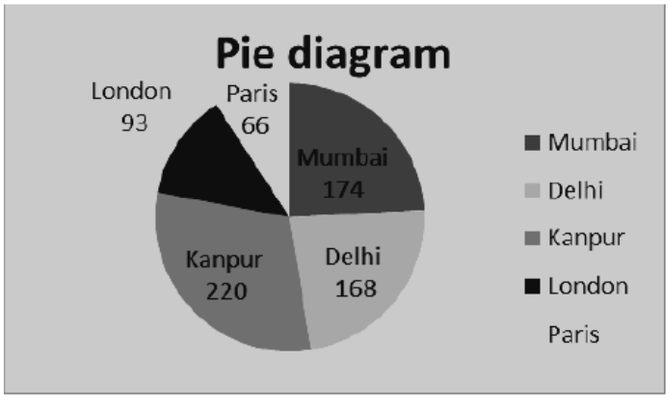
A pie chart is a type of graph that represents the data in the circular graph. The slices of pie show the relative size of the data, and it is a type of pictorial representation of data. A pie chart requires a list of categorical variables and numerical variables. Here, the term "pie" represents the whole, and the "slices" represent the parts of the whole.

Example 10: Represent the following data by a pie diagram.

Cities	Mumbai	Delhi	Kanpur	London	Paris
Infant Mortality	174	168	220	93	66

Solution: These values are converted into corresponding degrees in the circle taking total infant mortality 721 as equal to 360º. The calculation is shown in the table.

Cities	Infant Mortality	Angles (in degrees)
Mumbai	174	$\frac{174}{721}\times 360 = 86.88 = 87^{o}$
Delhi	168	$\frac{168}{721}\times 360 = 83.88 = 84^{o}$
Kanpur	220	$\frac{220}{721}\times 360 = 109.84 = 110^{o}$
London	93	$\frac{93}{721}\times 360 = 46.43 = 46^{o}$
Paris	66	$\frac{66}{721}\times 360 = 32.95 = 33^{o}$
Total	**712**	**360º**



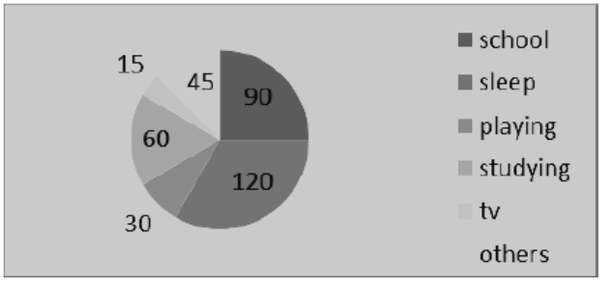
Example 11: The following table shows the numbers of hours spent by a child on different events on a working day.

Activity	No. of Hours
School	6
Sleep	8
Playing	2
Study	4
T. V.	1
Others	3

Represent the adjoining information on a pie chart.

Solution: The central angles for various observations can be calculated as:

Activity	No. of Hours	Measure of Central Angle
School	6	$(6/24 \times 360)° = 90°$
Sleep	8	$(8/24 \times 360)° = 120°$
Playing	2	$(2/24 \times 360)° = 30°$
Study	4	$(4/24 \times 360)° = 60°$
T. V.	1	$(1/24 \times 360)° = 15°$
Others	3	$(3/24 \times 360)° = 45°$



Notes

3.5 Cumulative Frequency and Ogives

Cumulative frequency curve or an ogive is the graphical representation of a cumulative frequency distribution.

There are two methods of constructing an ogive.

(i) Less than method (ii) More than method

3.5.1 Less than Method

Step I: If the frequency is in inclusive form convert it into exclusive form.

Step II: Construct a cumulative frequency table.

Step III: Mark upper class limits along x-axis.

Step IV: Mark the corresponding cumulative frequency along y-axis.

Step V: Plot the points and join them by a free hand curve.

Step VI: The lower limit of the first class interval becomes the upper limit of the imagined class with frequency 0. Join the imagined point (lower limit of first, 0) with the first point of the curve and so on.

Now, we get the required curve called an ogive.

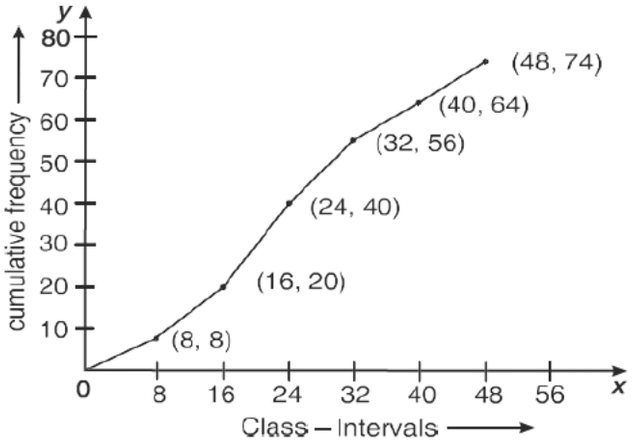
Example 12: Draw a cumulative frequency curve (or an ogive) for the following data.

Class Interval	0 – 8	8 – 16	16 – 24	24 – 32	32 – 40	40 – 48
Frequency	8	12	20	16	8	10

Solution: We first prepare the cumulative frequency distribution table.

Class-Interval	Frequency	Cumulative Frequency (Less than Type)
0 – 8	8	8
8 – 16	12	20
16 – 24	20	40
24 – 32	16	56
32 – 40	8	64
40 – 48	10	74

Notes



Mark the upper class limits along x-axis and the cumulative frequency along y-axis. Thus we plot the points (8, 8), (16, 20), (24, 40), (32, 56), (40, 64) and (48, 74). Join these points by a free hand curve. Complete the curve by joining the first point of the curve to the point (lower limit, 0).

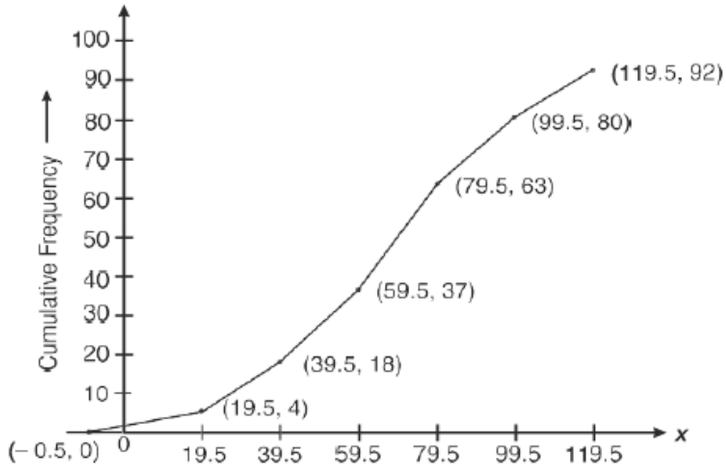
Example 13: Draw a cumulative frequency curve (or an ogive) for the following data:

Class Interval	0 – 19	20 – 39	40 – 59	60 – 79	80 – 99	100–119
Frequency	4	14	19	26	17	12

Solution: We first convert the class limits into true class limits and frequency distribution is converted into cumulative frequency distribution. Consider one imaginary point (lower limit of first class, 0).

Class Interval	Frequency	True Class Limits	Cumulative Frequency
0 – 19	4	–0.5 – 19.5	4
20 – 39	14	19.5 – 39.5	18
40 – 59	19	39.5 – 59.5	37
60 – 79	26	59.5 – 79.5	63
80 – 99	17	79.5 – 99.5	80
100 – 119	12	99.5 – 119.5	92

Notes



3.5.2 More than Method

To construct an ogive by more than type method, we apply the following steps:

Step I: Convert the frequency distribution into more than type cumulative frequency distribution by subtracting the frequency of each class from total frequency.

Step II: Mark the lower class limits on x-axis.

Step III: Mark the corresponding cumulative frequency on y-axis.

Step IV: Plot the points and join them by a free hand curve.

Example 14: The frequency distribution of scores obtained by 230 candidates in a medical entrance test is as follows:

Scores	400–450	450–500	500–550	550–600	600–650	650–700	700–750	750–800
Number of Candidates	20	35	40	32	24	27	18	34

Draw cumulative frequency curve by more than method.

Solution: More than method: Let us first prepare the cumulative frequency table by more than method as given below:

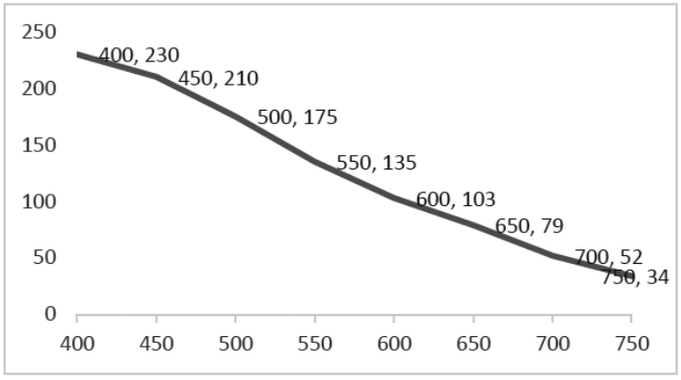
Scores	Number of Candidates	Scores More Than	Cumulative Frequency
400 – 450	20	400	230
450 – 500	35	450	210

Notes

Scores	Number of Candidates	Scores More Than	Cumulative Frequency
500 – 550	40	500	175
550 – 600	32	550	135
600 – 650	24	600	103
650 – 700	27	650	79
700 – 750	18	700	52
750 – 800	34	750	34

Mark the lower class limits on x-axis and cumulative frequency along y-axis.

Plot the points (400, 230), (450, 210), (500, 175), (550, 135), (600, 103), (650, 79), (700, 61), (750, 27) and an imagined point (800, 0)



3.6 Miscellaneous Questions

Example 15: An advertising company kept an account of response letters received each day over a period of 50 days. The observations were:

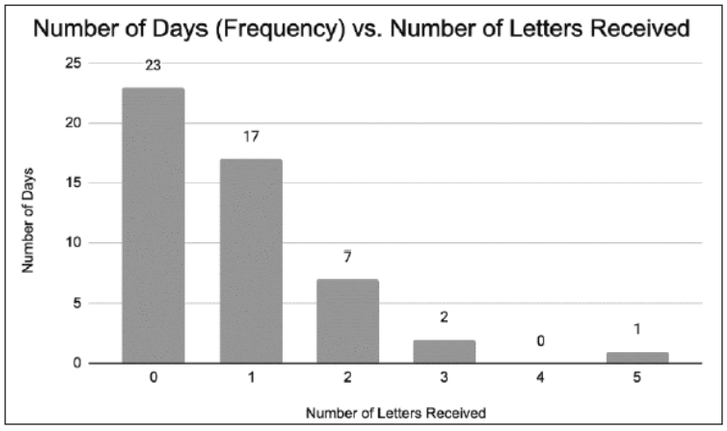
0	2	1	1	1	2	0	0	1	0	1	0	1	0	0	1	0	1	0	1	1	0	0	1	1	0
2	0	0	2	0	1	1	0	1	3	1	0	1	1	0	1	1	0	1	0	1	0	1	0		
2	5	1	1	2	0	0	0	0	5	0	1	2	0												

Construct a frequency table and draw a bar diagram to present the data.

Solution: The frequency distribution of letters received are shown in Table below. Figure depicts a frequency bar diagram for the number of letters received during a period of 50 days presented in Table below:

Notes

Number of Letters Received	Tally	Number of Days (Frequency)
0	𝍸 𝍸 𝍸 𝍸 |||	23
1	𝍸 𝍸 𝍸 ||	17
2	𝍸 ||	7
3	||	2
4		0
5 |	|	1
		Total = 50



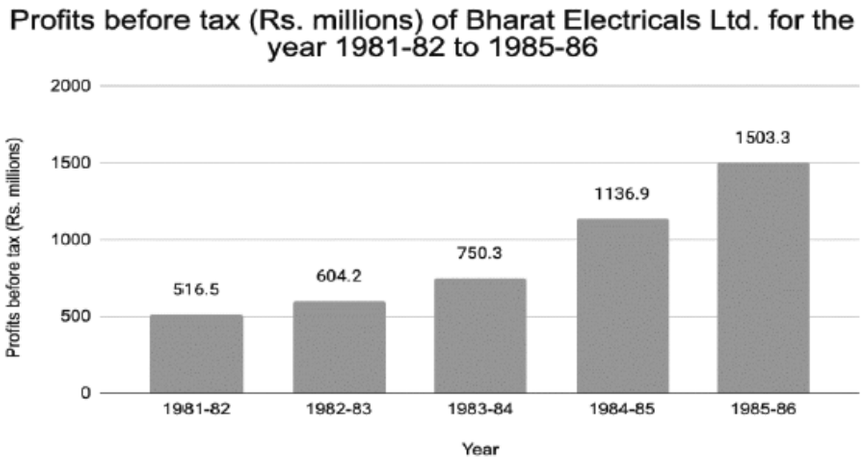
Example 16: The profits before tax of Bharat Heavy Electricals Ltd. are given below:

Year	Profits before Tax (Rs. Millions)
1981-82	516.5
1982-83	604.2
1983-84	750.3
1984-85	1136.9
1985-86	1503.3

Represent the data by a bar diagram.

Solution: The above data can be represented by a simple bar diagram:

Notes



Example 17: The following figures relate to the cost of construction of a house in Delhi:

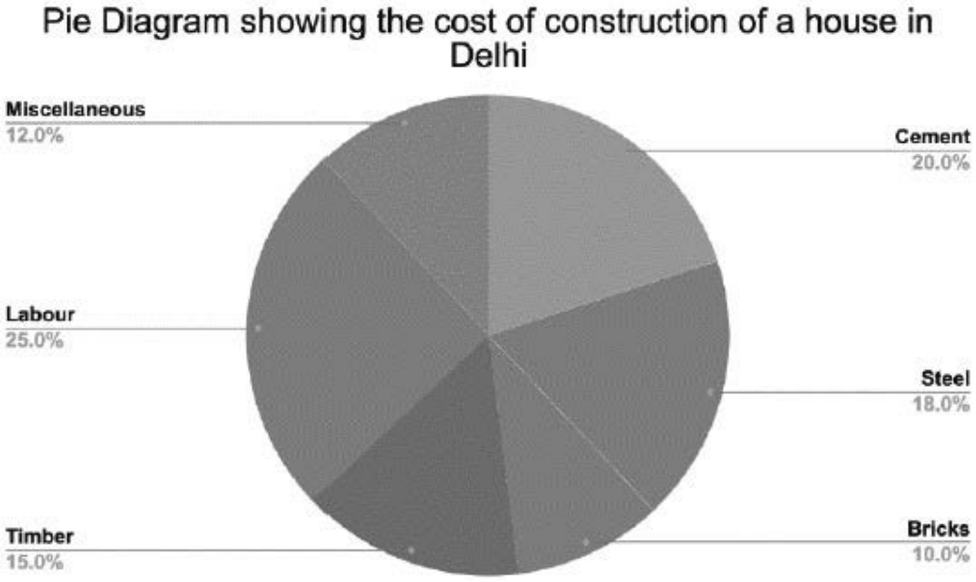
Items	Expenditure
Cement	20 %
Steel	18 %
Bricks	10 %
Timber	15 %
Labour	25 %
Miscellaneous	12 %

Represent the data by a suitable diagram.

Solution:

Items	Expenditure
Cement	20 × 3.6 = 72
Steel	18 × 3.6 = 64.8
Bricks	10 × 3.6 = 36
Timber	15 × 3.6 = 54
Labour	25 × 3.6 = 90
Miscellaneous	12 × 3.6 = 43.2

Notes



Example 18: Represent the following data by means of a histogram:

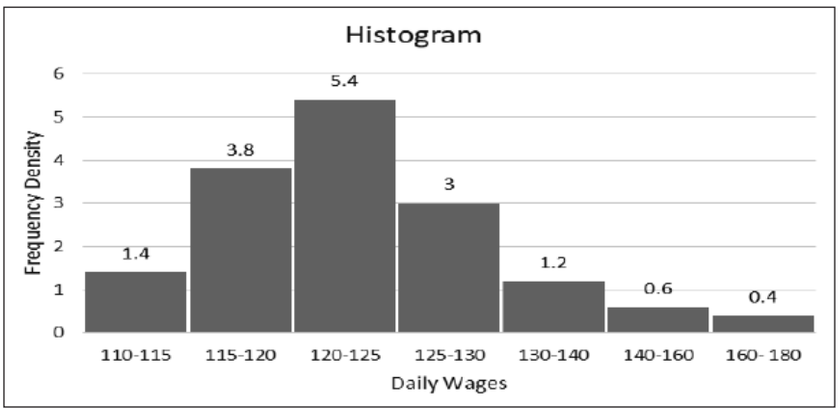
Daily Wages (in Rupees)	Number of Workers
110-115	7
115-120	19
120-125	27
125-130	15
130-140	12
140-160	12
160-180	8

Solution: In the given data, the wage intervals are not of equal width. In a histogram, each bar should represent an interval of the same width to accurately compare frequencies. The first few intervals (110-115, 115-120, 120-125, 125-130) have a width of 5 rupees. The subsequent intervals (130-140, 140-160, 160-180) have widths of 10, 20, and 20 rupees, respectively. If you plot this data directly, the bars corresponding to wider intervals will naturally be taller or have a larger area, even if the frequency is lower. This distorts the visual representation, making it harder to compare the data accurately.

To make histogram for unequal class intervals, first calculate the frequency densities by dividing the frequencies with their respective class width. Then plot frequency densities on y-axis and class interval on x-axis.

Notes

Daily Wages (in Rupees)	Number of Workers	Frequency Density
110-115	7	7/5 = 1.4
115-120	19	19/5 = 3.8
120-125	27	27/5 = 5.4
125-130	15	15/5 = 3
130-140	12	12/10 = 1.2
140-160	12	12/20 = 0.6
160- 180	8	8/20 = 0.4



Example 19: Represent the following frequency distribution by means of a histogram and superimpose thereon the corresponding frequency polygon and frequency curve:

Monthly Salary (in Rupees)	Number of Employees
3000 – 4000	20
4000 – 5000	30
5000 – 6000	60
6000 – 7000	75
7000 – 8000	115
8000 – 9000	100
9000 – 10000	60
10000 – 12000	80

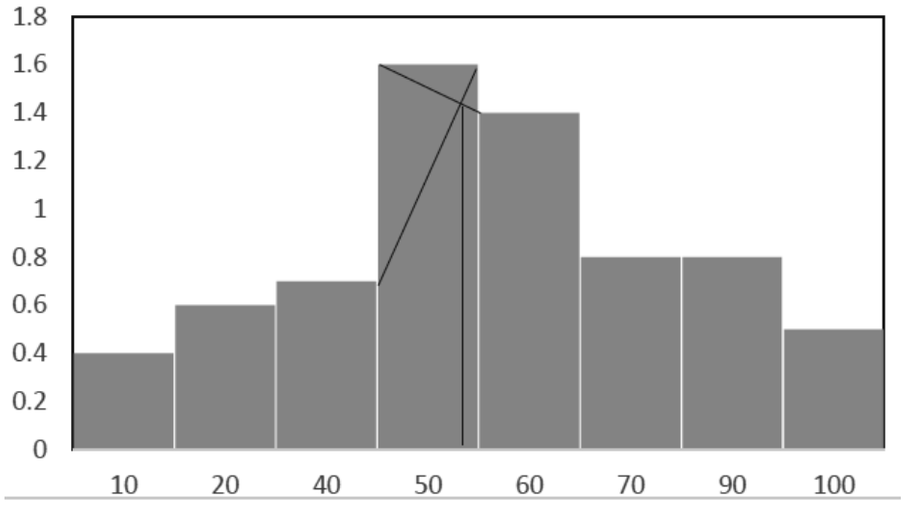
Notes

Example 20: Prepare a histogram from the following data and find the mode.

Marks	No. of Students
0-10	4
10-20	6
20-40	14
40-50	16
50-60	14
60-70	8
70-90	16
90-100	5

Solution: Since the class intervals are unequal, we first find frequency densities. The histogram is then plotted using frequency densities.

Class Intervals		Class Width	No. of Students	Frequency Density
0	10	10	4	0.4
10	20	10	6	0.6
20	40	20	14	0.7
40	50	10	16	1.6
50	60	10	14	1.4
60	70	10	8	0.8
70	90	20	16	0.8
90	100	10	5	0.5



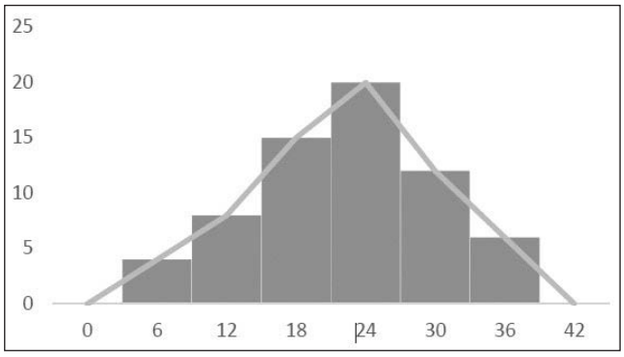
Notes

Example 21: Prepare a histogram and a frequency polygon from the following data.

Class Interval	0-6	6-12	12-18	18-24	24-30	30-36
Frequency	4	8	15	20	12	6

Solution:

Class Interval		**Frequency**
-6	0	0
0	6	4
6	12	8
12	18	15
18	24	20
24	30	12
30	36	6
36	42	0



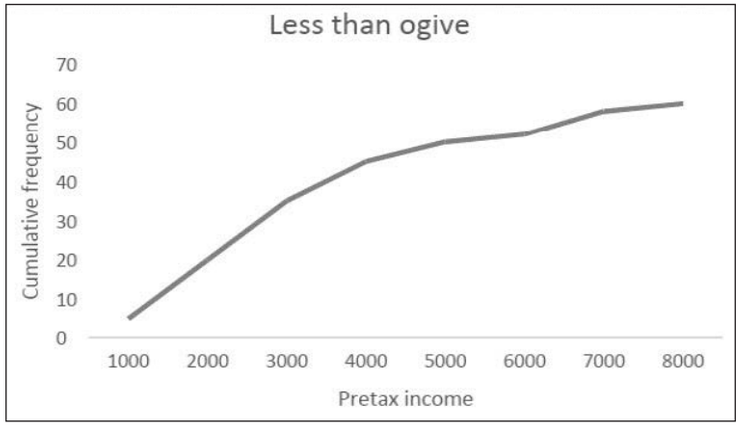
Example 22: Given below is the pretax monthly income of residents of an Industrial Town. Draw a less than Ogive.

Pretax Income (In Rupees)	**Number of Residence (in 1000)**
More than 7000	2
More than 6000	8
More than 5000	10
More than 4000	15
More than 3000	25
More than 2000	40
More than 1000	55
More than 0	60

Notes

Solution:

Pretax Income		Number of Residence (f)	Less than Cumulative Frequency
0	1000	5	5
1000	2000	15	20
2000	3000	15	35
3000	4000	10	45
4000	5000	5	50
5000	6000	2	52
6000	7000	6	58
7000	8000	2	60



3.7 Exercise

1. A family with monthly income of Rs. 20,000 had planned the following expenditure per month under various heads:

Heads	Expenditure (in Rs. 1000)
Grocery	4
Rent	5
Education of children	5
Medicine	2
Fuel	2
Entertainment	1
Miscellaneous	1

Draw the graph for the above data.

Notes

2. The following table gives the marks scored by 100 students in an entrance examination:

Marks	No. of Students (Frequency)
0–10	4
10–20	10
20–30	16
30–40	22
40–50	20
50–60	18
60–70	8
70–80	2

3. The following table shows the number of illiterate persons in the age group (10 – 58 years) in a town:

Age Group (in years)	Number of Illiterate Persons
10–16	175
17–23	325
24–30	100
31–37	150
38–44	250
45–51	400
52–58	525

Draw a histogram to represent the above data.

4. Draw a histogram to represent the following data which shows the monthly cost of living index at a city in a period of two years.

Cost of Living Index	Number of Months
440–460	2
460–480	4
480–500	3
500–520	5
520–540	3
540–560	2
560–580	1
580–600	4

Notes

5. Draw the histogram and frequency polygon of the following frequency distribution of the monthly wages:

Monthly Wages (in rupees)	Number of Workers
325 – 350	30
350 – 375	45
375 – 400	75
400 – 425	60
425 – 450	55
Total	**245**

6. Draw the frequency polygon representing the following frequency distribution.

Class Interval	Frequency
30 – 34	12
35 – 39	16
40 – 44	20
45 – 49	8
50 – 54	10
55 – 59	4

7. Construct a frequency polygon from the following data.

Score	Frequency
32 – 34	13
35 – 37	10
38 – 40	20
41 – 43	16
44 – 46	12
47 – 49	8

8. Draw an ogive for the following frequency distribution of less than method.

Marks	Number of Students
0 – 10	7
10 – 20	10
20 – 30	23
30 – 40	51
40 – 50	6
50 – 60	3

Notes

9. Represent the following data by an ogive by more than method.

Group	Frequency
0 – 10	4
10 – 20	4
20 – 30	7
30 – 40	10
40 – 50	12
50 – 60	8
60 – 70	5

10. Draw a cumulative frequency curve for the following frequency distribution.

Class Interval	Frequency
0 – 9	5
10 – 19	15
20 – 29	20
30 – 39	23
40 – 49	17
50 – 59	11
60 – 69	9

11. Draw an ogive (by more than method) to represent the following data, showing the monthly cost of living index of a city.

Cost of Living Index	No. of Months in a Period
340 – 350	10
350 – 360	19
360 – 370	24
370 – 380	18
380 – 390	16
390 – 400	13

12. The data shoes market share (in percentage) close by revenue of the following companies in a particular year:

Notes

Company	Market Share (in percentage)
BPL Ltd.	30
Hutchison	26
Bharti Telecom	19
Modi Telecom	12
Pacific India	5
Reliance	3
ParleG	2
Srinivas	2
Shyam Enterprises	1

Draw a pie diagram for the above data.

13. Which of the charts would you prefer to represent the following data pertaining to the monthly income of two families and the expenditure incurred by them?

Expenditure on	Family A (Income Rs. 17,000)	Family B (Income Rs. 10,000)
Food	4000	5400
Clothing	2800	3600
House Rent	3000	3500
Education	2300	2800
Miscellaneous	3000	5000
Saving Or Deficits	+1900	–300

14. The following data represent the outlays (in Rs. crore) bracket close by heads of development.

Heads of Development	Centre	States
Agriculture	4765	7039
Irrigation and Flood Control	6635	11395
Energy	9995	8293
Industry And Minerals	12770	2985
Transport And Communication	12200	5120
Social Services	8216	1420
Total	**54581**	**36252**

Notes

15. Draw a histogram and frequency polygon from the following data:

Class	Frequency
0-10	4
10-20	6
20-40	14
40-50	16
50-60	14
60-70	8
70-90	16
90-100	5

16. Construct a histogram from the following data:

Class Limits	Frequency
9-10	16
10-11	22
11-12	45
12-13	60
13-14	50
14-15	24
15-16	10

17. The frequency distribution of marks obtained by 60 students of a class in a college are given below:

Marks	30-34	35-39	40-44	45-49	50-54	55-59	60-64
No. of Students	3	5	12	18	14	6	2

Draw a histogram for this distribution and find modal value. Draw a cumulative frequency curve also.

18. Draw a less than Ogive from the following data:

Weekly Income (Rs.) (Equal to or More Than)	No. of Families
12,000	0
11,000	6
10,000	14
8,000	26
6,000	42
4,000	54

Notes

Weekly Income (Rs.) (Equal to or More Than)	No. of Families
3,000	62
2,000	70
1,000	80

From the graph estimate the number of families in the income range of Rs. 24,000 and Rs. 10,500. Also, find the maximum income of the lowest of 25% of the families.

UNIT - II

LESSON

4

Measures of Central Tendency

STRUCTURE

4.1 Measures of Central Tendency

An average is a value which is representative of a set of data. Average value may also be termed as measure of central tendency. There are three types of common measures of central tendency.

(a) Arithmetic mean or average

(b) Median

(c) Mode

Properties for an Ideal Measure of Central Tendency

- It should be rigidly defined.
- It should be readily comprehensible and easy to calculate.
- It should be based on all the observations.
- It should be suitable for further mathematical treatments.
- It should not be affected much by fluctuations of sampling.
- It should not be affected much by extreme values.

Notes

4.2 Arithmetic Mean

If $x_1, x_2, x_3, \ldots, x_n$ are n numbers, then their arithmetic mean (A.M.) is defined by.

Arithmetic Mean = $\bar{x} = (x_1 + x_2 + x_3 + \cdots + x_n)/n$

where,

x_1, x_2, x_3, x_n are the observations.

n is the number of observations.

Alternatively, one can symbolically write it as shown below:

Arithmetic Mean Formula = $\bar{x} = \sum x_i f_i / \sum f_i$

In the above equation, the symbol $\sum$ known as sigma. It implies summation of the values.

4.2.1 Merits and Demerits of Arithmetic Mean

Merits:

1. Rigidly defined
2. Easy to understand and calculate
3. Based upon all the observations.
4. Suitable for further mathematical calculations
5. It is least affected by fluctuations of sampling.

Arithmetic mean is sometimes referred as stable average.

Demerits:

1. It cannot be obtained by inspection nor it can be located graphically.
2. A.M. cannot be used if we are dealing with qualitative characteristic which cannot be measured quantitatively. In this case only Median is used.
3. Arithmetic mean cannot be obtained if a single observation is missing or lost or is illegible unless we drop it out and compute the arithmetic mean of the remaining values.
4. It is affected very much by extreme values (in this case it gives a distorted pictures of the distance and does not represent the distance.)
5. It may lead to wrong conclusions if the details of the data from which it is computed are not given.

Notes

6. It cannot be calculated if the extreme class is open.
7. In extremely asymmetrical (skewed) distance, arithmetic mean is not a suitable measure of location.

4.2.2 Calculation of Arithmetic Mean

(a) Direct Method

Direct method is the most basic way to calculate the mean of grouped data. The procedures for utilizing the direct technique to obtain the mean for grouped data are outlined below:

Make a table with four columns, as shown below:

Column 1 – Class Intervals

Column 2 – Corresponding class marks, marked by x_i. (In case of discrete data, where class intervals are not given, then value of the data forms the x_i)

Column 3 – Frequencies (f_i) in the corresponding class

Column 4 – $x_i f_i$ (corresponding product of column 2 and column 3)

Determine the mean using the Formula = $\sum x_i f_i / \sum f_i$

Example 1: Find the mean of 20, 22, 25, 28, 30.

Solution:

A.M. = (20 + 22 + 25 + 28 + 30)/5

= 125/5 = 25

Example 2: Find the arithmetic mean of the following distribution:

Number (x)	8	10	15	20
Frequency (f)	5	8	8	4
fx	40	80	120	80

Solution:

$\Sigma fx = 40 + 80 + 120 + 80 = 320$

$f = 5 + 8 + 8 + 4 = 25$

A.M. = $\Sigma fx/f = 320/25 = 12.8$

(b) Shortcut Method/Assumed Mean Method

When the direct technique becomes too time-consuming, we use the assumed mean method to calculate the average of a set of grouped data.

Notes

We can use the assumed mean method to detect and calculate the mean by following the procedures listed below:

Make a table using the five columns listed below:

Column 1: Class Intervals between classes.

Column 2: Class marks, represented by x_i. Assumed Mean A: Pick the middle value from the class marks and indicate it as A.

Column 3: Determine the relevant deviations using the formula

$$d_i = x_i - A.$$

Column 4: Frequencies (f_i) of the corresponding class

Column 5: Mean of deviated values = $\sum f_i d_i / \sum f_i$

Finally, calculate the Mean of the original data by adding the assumed mean to the average of the deviated values.

We usually assume a value as the average in this manner (namely, A). This value is used to calculate the deviations that the formula is based on. In addition, the information will be presented as a frequency distribution table with classifications. As a result, the formula for calculating the mean using the assumed mean technique is:

$$\text{Mean } (\bar{x}) = A + \sum f_i d_i / \sum f_i$$

Example 3: Find the arithmetic mean of the following distribution:

Class	0–10	10–20	20–30	30–40	40–50
Frequency	7	8	20	10	5

Solution: Let assumed mean (A) = 25.

Class	Mid-value x	Frequency f	x – 25 = d	fd
0–10	5	7	–20	–140
10–20	15	8	–10	–80
20–30	25	20	0	0
30–40	35	10	+10	+100
40–50	45	5	+20	+100
Total		Σf = 50		Σfd = –20

A.M. = 25 + [(–20)/50]

= 24.6

(c) Step-Deviation Method

The shift of origin and scale method are other name for step deviation. When calculating the mean for grouped data, we use the step deviation

Notes

approach to simplify the calculations. The following are the steps to take while using the step deviation method:

Make a table with five columns, as shown below:

Column 1: Class intervals.

Column 2: Corresponding class marks, represented by x_i. Take the middle value from the class marks and indicate it as A.

Column 3: Corresponding frequencies (f_i) in the next column.

Column 4: Determine the corresponding deviations using the formula $d_i = x_i - A$. Use the formula $u_i = d_i/h$ to calculate the values of u_i, where h is the class width.

Column 5: Multiply the corresponding frequencies (f_i) with u_i in the next column.

The step-deviation method can be used to find the mean when the data values are large. The formula is as follows:

$$\text{Mean } (\bar{x}) = A + (h \sum f_i u_i / \sum f_i)$$

Example 4: Find the arithmetic mean of the data given in example 3 by step deviation method

Solution: Let a = 25

Class	Mid-value (x)	Frequency (f)	$u = \frac{x - a}{h}$	fu
0–10	5	7	–2	–14
10–20	15	8	–1	–8
20–30	25	20	0	0
30–40	35	10	+1	+10
40–50	45	5	+2	+10
Total		$\Sigma f = 50$		$\Sigma fu = -2$

A.M. = 25 + (10*(_2)/50)

= 24.6

4.3 Median

Median is defined as the measure of the central unit when they are arranged in ascending or descending order of magnitude.

Notes

Median = ((n + 1)/2)th term if given data set has odd number of values

= average of (n/2)th and ((n/2)+1)th observation if data set is even.

Steps for calculating the Median

Step 1: Sort your observations into ascending or descending order.

Step 2: The median is the middle observation if the number of observations is odd, or the average of the two middle observations if the number of observations is even.

4.3.1 Merits and Demerits of Median

Merits:

1. It is rigidly defined.
2. Easily calculated and understand. In some cases, it can be located merely by inspections.
3. It is not at all affected by extreme values.
4. It can be calculated for distributions with open-end classes.
5. When data is qualitative in nature, median is the only way to find average.

Demerits:

1. In case of even number of observations, it is not obtained exactly but is estimated by taking the mean of two middle values.
2. It is not based on all observations. (It is insensitive)

4.3.2 Determining Median graphically

Median can be located graphically by any of the two ways:

1. Draw two ogives. Now find the point of intersection of the two ogives. From the point of intersection, draw a perpendicular on x-axis. The point at which the perpendicular touches the x-axis gives the median.
2. Draw only one ogive by 'less than method', taking the variable on x-axis and frequency on y-axis. Find median = size of (n/2)th item. Locate this value on y-axis. From this point draw a perpendicular on the less than ogive curve, parallel to x-axis. From the point where it meets the ogive, draw another perpendicular on the x-axis.

Notes

The point at which the perpendicular touches the x-axis gives the median.

Note: We can determine the other partition values like quartiles, deciles, percentiles, etc. by using method 2 described above.

Example 5: Find the median of 6, 8, 9, 10, 11, 12, 13.

Solution: Total number of items = 7

The middle item = ((7 + 1)/2) = 4

Median = Value of the 4th unit = 10

Calculation of Median for Grouped Data

$$Median = l + \frac{\left(\frac{n}{2}\right) - cf}{f} \times h$$

where, l is the lower limit of the median class,

f is the frequency of the class,

h is the width of the median class

cf is the cumulative frequency of the class preceding the median-class and

n is total frequency of the data.

Example 6: Find the value of Median from the following data:

No. of Days for which Absent (less than)	5	10	15	20	25	30	35	40	45
No. of students	29	224	465	582	634	644	650	653	655

Solution: The given cumulative frequency distribution will first be converted into ordinary frequency as under:

Class-Interval	Cumulative Frequency	Ordinary Frequency
0–5	29	29=29
5–10	224	224–29=195
10–15	465	465–224=241
15–20	582	582–465=117
20–25	634	634–582=52
25–30	644	644–634=10
30–35	650	650–644=6
35–40	653	653–650=3
40–45	655	655–653=2

Notes

Median = 655/2 = 327.5th item

327.5th item lies in 10-15 which is the median class.

$$median = 10 + \frac{327.5 - 224}{241} \times 5$$

$$= 12.147$$

4.4 Mode

The mode is that value of the observation in the series which occurs the largest number of times or which has greatest frequency. Thus, mode is the most popular item of the series around which there is the highest frequency density. Usually denoted by M_o.

When we speak of 'average student', 'average collar size', 'average T-shirt size', 'average shoe size', we are referring to mode.

Mode can be calculated for ungrouped and grouped series. A distribution may have more than one mode.

For Ungrouped Data

The mode is calculated just by inspection or by counting the number of items. It is the value of the observation corresponding to maximum frequency.

Example 7: Find the mode of the following items: 0, 1, 6, 7, 2, 3, 7, 6, 6, 2, 6, 0, 5, 6, 0.

Solution: As 6 occurs 5 times and no other item occurs 5 or more than 5 times, hence the mode is 6.

For Grouped Data

To calculate the mode, we first ensure that it is continuous exclusive series having equal class intervals. By inspection locate the modal class. The modal class is one having maximum frequency. The mode is then given by

$$M_o = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

where l is the lower limit of the modal class, f_0 is the frequency of the modal class, h is the width of the class, f_1 is the frequency before the modal class and f_2 is the frequency after the modal class.

Notes

Note:

1. In moderately skewed or asymmetrical distribution, when the values of any two averages is given then the third average can be obtained using the following empirical formula:

   Empirical formula = Mean – Mode = 3 (Mean – Median)

2. Mode is used when the most typical value of a distribution is desired. For example, mode is used to find average amount of money spend by students per month, average collar size, average shoe size, etc.

4.4.1 Merits and Demerits of Mode

Merits

1. It is simple and easy to understand.
2. In discrete series, mode can be located even by inspection.
3. It is not unduly affected by extreme values.
4. It can be determined graphically.
5. It can be used to describe qualitative phenomenon.

Demerits

1. Mode is sometimes ill-defined. The value of mode cannot be determined always.
2. It is not based on all the observations of the series.
3. It is not capable of further mathematical treatment.
4. In some cases, there exist more than one mode.

4.4.2 Locating mode graphically

The value of the mode can be graphically determined in a frequency distribution. The steps for locating the mode graphically are:

1. Draw a histogram of the given data.
2. The bar with the highest peak is the modal class bar. Now draw two lines diagonally inside the modal class bar starting from each upper corner of the bar to the upper corner of the adjacent bars.
3. Draw a perpendicular from the intersection of these lines to the x-axis. The point at which the perpendicular touches the x-axis gives the modal value.

Notes

Example 8: The heights, in cm, of 50 students are recorded. Calculate mode.

Height (in cm)	125-130	130-135	135-140	140-145	145-150
Number of Students	7	14	10	10	9

Solution:

Here, the maximum frequency is 14 and the corresponding class is 130-135. So, 130-135 is the modal class such that

$$l = 130,\ h = 5,\ f_0 = 14,\ f_1 = 7,\ f_2 = 10$$

$$\text{mode} = 130 + [14 - 7/(14 - 7) + (14 - 10)]*5 = 133.18$$

Hence, the modal height = 133.18

4.5 Weighted Mean

In case of arithmetic mean, we suppose that all the items in the distribution have equal importance. This is not actually so. Some of the items in a distribution may be more important than others. Then in such cases, proper weightage is given to various items. The weights attached to each item is proportional to the importance of the item in the distribution.

Let w_i be the weight attached to the item x_i, i = 1, 2,..... n

then,

Weighted arithmetic mean or weighted mean is = $\Sigma w_i\ x_i/\Sigma w_i$

Weighted mean gives the result equal to the simple mean if the weights assigned to each of the items are equal. If smaller weights are given to smaller items and larger weights to larger items then it results in higher value than simple mean. If the weights attached to larger items are smaller and those attached to smaller items are larger than it results in smaller value than the simple mean.

4.6 Partition Values

Partitions values are positional measures just like median and are those values of the variate which divide the series into a number of equal parts. The most common partition values besides median are quartiles, deciles and percentiles which divide the series into four, ten and hundred equal

Notes

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

4.6.1 Quartiles

The values of a variate which divide the series into four equal parts are called quartiles. First arrange the data in ascending or descending order. We know that three points are required to divide the data into four equal parts, so we have three quartiles, denoted by Q_1, Q_2 and Q_3.

The first quartile (Q_1), also known as lower quartile is that value of the variate below which there are 25% of the observations and above which there are 75% of the observations.

The second quartile (Q_2), also known as middle quartile or median is that value of the variate which divides the series into two equal parts i.e. 50% of the observations are below it 50% of the observations above it.

The third quartile (Q_3), also known as upper quartile is that value of the variate below which there are 75% of the observations and above which there are 25% of the observations.

We see that $Q_1 < Q_2 < Q_3$

Computation of Quartiles

In case of individual series or discrete series

First arrange the data into ascending or descending order of magnitude. Now

Q_1 = value of $\left(\frac{n+1}{4}\right)^{th}$ item in the series

Q_2 = value of $\left(\frac{2(n+1)}{4}\right)^{th}$ item in the series

Q_3 = value of $\left(\frac{3(n+1)}{4}\right)^{th}$ item in the series

In case of continuous series

We first identify the quartile class. The computation of quartiles in case of grouped data is done exactly in the same manner as done in the case of median. The formula for calculating quartiles are:

Notes

$$Q_1 = l + \frac{\left(\frac{n}{4}\right) - cf}{f} \times h$$

$$Q_2 = l + \frac{\left(\frac{2n}{4}\right) - cf}{f} \times h$$

$$Q_3 = l + \frac{\left(\frac{3n}{4}\right) - cf}{f} \times h$$

where,

l = lower limit of the quartile class

cf = cumulative frequency of the class prior to the quartile class

f = frequency of the quartile class

h = width of the quartile class

n = total number of observations in the distribution

Note: Q_2 is same as the median

4.6.2 Deciles

The values of a variate that divides the series into ten equal parts are called deciles. Since nine points are required to divide the arranged data into ten equal parts, there are 9 deciles denoted by D_1, D_2,….., D_9. Each part contains 10% of the data.

Computation of deciles in case of individual series or discrete series

The general formula of jth decile is

$$D_j = \text{value of } \left(\frac{j(n+1)}{10}\right)^{th} \text{ item in the series}$$

where j = 1, 2,…., 9

Notes

In case of grouped series

The formula for jth decile is

$$D_j = l + \frac{\left(\frac{jn}{10}\right) - cf}{f} \times h$$

where j = 1, 2,…., 9

where cf is the cumulative frequency preceding the jth decile class. Other notations have usual meaning.

4.6.3 Percentile

The value of the variate which divide the series into 100 equal parts are called percentiles. Since ninety nine points are required to divide the data into 100 equal parts, there are 99 percentile values denoted by Pj (j = 1, 2,…..,99). Each percentile contains 1% of the total number of observations.

Computation of percentiles

In case of discrete or individual series

The formula for jth percentile is

$$\text{Pj} = \text{value of } \left(\frac{j(n+1)}{100}\right)^{th} \text{ item in the series}$$

where j = 1, 2,…., 9

In case of continuous series

The formula for the jth percentile is

$$P_j = l + \frac{\left(\frac{jn}{100}\right) - cf}{f} \times h$$

where j = 1, 2,…., 99

cf is the cumulative frequency preceding the jth percentile class.

Notes

Example 9: The data set given below has 19 observations, calculate all three quartiles.

Solution: First of all, we need to sort this data in the ascending order as given below;

Here total observation count N = 19, now we will calculate the quartiles using their respective formulas given above in this section.

Q_1 = 1 * (19+1)/4 = 20/4 = 5 (i.e. 5th observation) = 33

Q_2 = 2 * (19+1)/4 = 40/4 = 10 (i.e. 10th observation) = 48

Q_3 = 3 * (19+1)/4 = 60/4 = 15 (i.e. 15th observation) = 61

Example 10: Below table shows the scores (out of 100) in a mathematics test for 30 students in a class. Calculate the 1st, 4th and 5th decile values for given data.

Roll No.	Test Score	Roll No.	Test Score
1	64	16	44
2	93	17	71
3	66	18	50
4	75	19	47
5	97	20	88
6	92	21	45
7	62	22	74
8	49	23	82
9	67	24	57
10	63	25	61
11	78	26	66
12	56	27	95
13	58	28	78
14	99	29	77
15	86	30	82

Notes

Solution:

The very first step that we have to do is to arrange the given data in the ascending order based on test scores as shown below:

Roll No.	Test Score	Roll No.	Test Score
16	44	17	71
21	45	22	74
19	47	4	75
8	49	29	77
18	50	11	78
12	56	28	78
24	57	23	82
13	58	30	82
25	61	15	86
7	62	20	88
10	63	6	92
1	64	2	93
3	66	27	95
26	66	5	97
9	67	14	99

Here total number of observations, N = 30

Now, using the decile formula discussed above in this section, deciles will be calculated as follows:

$D_1 = 1\times (30+1)/10 = 31/10 = 3.1$

= 3rd observation + 0.1× (4th observation – 3rd observation)

= 47 + 0.1× (49-47) = 47.2

$D_4 = 4\times (30+1)/10 = 124/10 = 12.4$

= 12th observation + 0.4× (13th observation – 12th observation)

= 64 + 0.4× (66-64) = 64.8

Notes

$D_5 = 5 \times (30 + 1)/10 = 155/10 = 15.5$

= 15th observation + 0.5 × (16th observation – 15th observation)

= 67 + 0.5 × (71 – 67) = 69

Example 11: The heights (in cm) of 50 first year students of a college are recorded and shared in the table below. Calculate the heights at 30th, 46th and the 90th percentile values from the given distribution.

153.5	154.6	155.5	154.9	150.2	152.1	154.7	150.4	155.2	154.1
157.5	158.1	161.4	160.3	155.7	156.3	159.4	155.8	161.2	157.7
164.7	166.6	168.5	167.8	161.5	164.1	167.5	162.2	167.9	165.2
171.5	174.1	177.8	175.5	169.3	171.4	174.8	170.2	176.6	172.4
181.4	182.7	185.4	184.6	179.2	180.5	183.5	179.7	184.9	182.6

Solution: First of all we have to arrange this data set in the ascending order of magnitude.

150.2	150.4	152.1	153.5	154.1	154.6	154.7	154.9	155.2	155.5
155.7	155.8	156.3	157.5	157.7	158.1	159.4	160.3	161.2	161.4
161.5	162.2	164.1	164.7	165.2	166.6	167.5	167.8	167.9	168.5
169.3	170.2	171.4	171.5	172.4	174.1	174.8	175.5	176.6	177.8
179.2	179.7	180.5	181.4	182.6	182.7	183.5	184.6	184.9	185.4

Here N = 50, now using the percentile formula we will calculate P_{30}, P_{46} & P_{90}

$P_{30} = 30 \times (50 + 1)/100 = 1530/100 = 15.3$

= 15th observation + 0.3 × (16th observation – 15th observation)

= 157.7 + 0.3 × (158.1 – 157.7) = 157.82

$P_{46} = 46 \times (50 + 1)/100 = 2346/100 = 23.46$

= 23rd observation + 0.46 × (24th observation – 23rd observation)

= 164.1 + 0.46 × (164.7 – 164.1) = 164.38

$P_{90} = 90 \times (50 + 1)/100 = 4590/100 = 45.9$

= 45th observation + 0.9 × (46th observation – 45th observation)

= 182.6 + 0.9 × (182.7 – 182.6) = 182.6

Therefore, the heights (in cm) at 30th, 46th, and 90th percentile values are 157.82, 164.38, and 182.69 respectively.

Notes

4.7 Miscellaneous Questions

Example 12: The following data give an actual distribution, obtained by tossing ten coins 1024 times and recording the number of heads that appeared on each toss. What is the average number of heads per toss?

Number of Heads	Frequency
0	1
1	16
2	42
3	126
4	199
5	253
6	209
7	118
8	53
9	4
10	3

Solution: Computation of average number of heads per toss.

Number of Heads (m)	Frequency (f)	mf
0	1	0
1	16	16
2	42	84
3	126	378
4	199	796
5	253	1265
6	209	1254
7	118	826
8	53	424
9	4	36
10	3	30
	n = 1024	Σmf = 5109

Arithmetic average $a = \frac{\Sigma mf}{n} = \frac{5109}{1024} = 5$ heads per toss approx.

Thus, the average number of heads per toss is 5.

Notes

Example 13: The following table gives the population of males at different age-groups of the U.K. and India at the time of a census.

Age-Group	U.K. (Lakhs)	India (Lakhs)
0-5	18	214
5-10	19	258
10-15	20	222
15-20	18	157
20-25	16	145
25-30	14	161
30-40	27	257
40-50	25	184
50-60	19	120
Above 60	17	100

Compare the average age of males in the two countries, and account for difference, if any.

Solution:

Age-group (m)	Mid-Values (x)	d= x-a = x-27.5	U.K.		India	
			Population of males in U.K. (f)	fd	Population of males in India (f)	fd
0-5	2.5	-25	18	-450	214	-5350
5-10	7.5	-20	19	-380	258	-5160
10-15	12.5	-15	20	-300	222	-3330
15-20	17.5	-10	18	-180	157	-1570
20-25	22.5	-5	16	-80	145	-725
25-30	27.5	0	14	0	161	0
30-40	35.0	7.5	27	202.5	257	1927
40-50	45.0	17.5	25	437.5	184	3220
50-60	55.0	27.5	19	522.5	120	3300
Above 60	65.0	37.5	17	637.5	100	3750
			n=193	Σfd = 410	n=1818	Σfd = -3937

Notes

The average age of males in the U.K. is calculated as follows:

$$\bar{x} = a + \frac{\Sigma fd}{n} = 27.5 + \frac{410}{193} = 27.5 + 2.12 = 29.62 \, years$$

The average age of males in the India is calculated as follows:

$$\bar{x} = a + \frac{\Sigma fd}{n} = 27.5 + \frac{-3937.5}{1818} = 27.5 - 2.17 = 25.33 \, years$$

Thus the average age of males in U.K. is higher than the average age of males in India.

Example 14: Find the average wage of a labourer from the following table:

Wage (in Rupees) Above	No. of Labourers
300	650
310	500
320	425
330	375
340	300
350	275
360	250
370	100

Solution: We have a cumulative frequency distribution table. To find the average wage, we need to convert this cumulative frequency distribution into a simple frequency distribution.

Step 1: Find the class intervals

To determine the class intervals, we need to find the difference between consecutive values in the "Wage (in Rupees) above" column. For example, 310 – 300 = 10, 320 – 310 = 10 and so on. So, the class interval is 10.

Step 2: Determine the frequency for each class

The frequency of the first class (300-310) is the total number of labourers (650) minus the number of labourers earning more than 310 (500).

The frequency of the second class (310-320) is the number of labourers earning more than 310 (500) minus the number of labourers earning more than 320 (425).

Notes

Wage (in Rupees) Above	No. of Labourers
300-310	650 - 500 = 150
310-320	500 - 425 = 75
320-330	425 - 375 = 50
330-340	375 - 300 = 75
340-350	300 - 275 = 25
350-360	275 - 250 = 25
360-370	250 - 100 = 150
370 and above	100

Step 3: Calculate the average wage

Now that we have a simple frequency distribution, we can calculate the average wage.

Wage (in Rupees) Above	No. of Labourers (f)	Mid-Point of the Classes (m)	x = m-345	dx=x/10	fdx
300-310	150	305	-40	-4	-600
310-320	75	315	-30	-3	-225
320-330	50	325	-20	-2	-100
330-340	75	335	-10	-1	-75
340-350	25	345	0	0	0
350-360	25	355	10	1	25
360-370	150	365	20	2	300
370 and above	100	375	30	3	300
	n=650				Σfdx= -375

The average wage of a labourer is calculated as follows:

$$\overline{x} = a + \frac{\Sigma f.d}{n} x_i = 345 + \frac{-375}{650} \times 10 = 345 + 5.77 = 339.23 \, rupees.$$

Example 15: The following table indicates the increase in cost of living over July 1972 for a working-class family as at 1st January 1975 and the weights assigned to various groups.

Notes

Group	Percentage Increase	Weights
Food	29	7.5
Rent	54	2.0
Clothing	97.5	1.5
Fuel and Light	75	1.0
Other items	75	0.5

Find out the weighted average of the increase in cost of living.

Solution: To find the weighted average of the increase in the cost of living, we need to multiply each percentage increase by its corresponding weight, sum up these products, and then divide the total by the sum of the weights.

Group	Percentage Increase (p)	Weights (w)	pw
Food	29	7.5	217.50
Rent	54	2.0	108.00
Clothing	97.5	1.5	146.25
Fuel and Light	75	1.0	75.00
Other items	75	0.5	37.50
		Σw = 12.5	Σpw = 584.25

$$\text{Weighted average} = \frac{\sum pw}{\sum w} = \frac{584.25}{12.5} = 46.74\%$$

Example 16: The arithmetic mean, the mode and the median of a group of 75 observations were calculated to be 27, 34 and 29 respectively. It was later discovered that one observation was wrongly read as 43 instead of the correct value 53. Examine to what extent the calculated values of the three averages will be affected by the discovery of this error.

Solution:

1. The arithmetic mean is calculated as the sum of all observations divided by the number of observations.

   Initially:

$$\bar{X} = \frac{\Sigma X}{N}$$

Notes

$\Sigma X = N \bar{X}$

$N = 75, \quad \bar{X} = 27$

Incorrect: $\Sigma X = 75 \times 27 = 2025$

Correct: $\Sigma X = 2025 - 43 + 53 = 2035$

Correct $\bar{X} = \frac{2035}{75} = 27.13$

2. **Mode:** The mode is the value that appears most frequently in the data set. Since correcting the observation from 43 to 53 changes just one value, and if neither 43 nor 53 was the mode, the mode remains unchanged at 34.
3. **Median:** The median is the middle value when the data set is ordered. Given there are 75 observations, the median is the 38th observation. If the incorrect value of 43 did not alter the central position (median), changing it to 53 would have no effect. However, if 43 was close to the median, this correction might slightly change the median, but it's unlikely given the median was 29, and this correction involves numbers larger than the median.

The discovery of the error has a minimal impact on the arithmetic mean and likely no impact on the mode or median.

Example 17: Find the weighted arithmetic mean of first n natural numbers who weights are equal to corresponding numbers.

Solution: To find the weighted arithmetic mean of the first n natural numbers where the weights are equal to the corresponding numbers, you can use the following approach:

The first n natural numbers are 1,2, 3…,n.

The weights for these numbers are also 1,2, 3…,n.

For the first n natural numbers,

1. The weighted sum is: $\sum(i.i) = \sum i^2$
2. The sum of squares of the first n natural numbers is given by:

$$\sum i^2 = n(n + 1)(2n + 1)/6$$

Calculate the Total Weight:

The total weight is: $\sum i = n(n + 1)/2$

Notes

Calculate the Weighted Mean:

$$\text{Weighted Mean} = \text{Weighted Sum/Total Weight}$$
$$= n(n + 1)(2n + 1)/6/n(n + 1)2$$
$$= 2n + 1/3$$

Hence, the weighted arithmetic mean of the first n natural numbers where the weights are equal to the corresponding numbers is 2n+1/3.

Example 18: Calculate the median from the data given below:

Class Interval	5-9	10-14	15-19	20-24	25-29	30-34	35-39
Frequency	8	15	18	30	16	12	6

Solution:

To find the median, we'll follow these steps:

1. **Calculate the cumulative frequency for each class interval:**
   - For 5-9: 8
   - For 10-14: 8+15=23
   - For 15-19: 23+18=41
   - For 20-24: 41+30=71
   - For 25-29: 71+16=87
   - For 30-34: 87+12=99
   - For 35-39: 99+6=105
2. **Identify the Median Class:** The total number of observations (N) is 105. Hence, Median position=N/2=105/2=52.5.
3. **Locate the median class by finding the cumulative frequency just greater than 52.5:** The cumulative frequency just greater than 52.5 is 71, which corresponds to the class interval **20-24**. So, the median class is **20-24**.
4. **Apply the Median Formula:**

The formula for the median in grouped data is:

$$Median = L + \frac{\left(\frac{N}{2}\right) - C}{f} \times h$$

Notes

where:

L = Lower boundary of the median class = 20

N = Total number of observations = 105

C = Cumulative frequency of the class before the median class = 41

f = Frequency of the median class = 30

h = Class width of the median class = 24 – 20 + 1 = 5

Substitute the values in the formula

$$Median = 20 + \frac{52.5 - 41}{30} \times 5 = 21.92$$

Example 19: Find the missing frequency from the following distribution if median is 35 and N = 170.

Variable	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	10	20	-	40	-	25	15

Solution:

To find the missing frequencies (say f_1 and f_2), we'll follow these steps:

1. **Determine the Median Class:**

   The median is given as 35, which falls in the class interval **30-40**.

2. **Calculate Cumulative Frequencies:**

   Calculate cumulative frequencies up to the median class:

   - Cumulative frequency up to 0-10: 10
   - Cumulative frequency up to 10-20: 10+20=30
   - Cumulative frequency up to 20-30: 30+f1
   - Cumulative frequency up to 30-40: 30+f1+40

Let's denote the cumulative frequency just before the median class by C. So, $C=30+f_1$.

1. **The Median Formula:**

$$Median = L + \frac{\left(\frac{N}{2}\right) - C}{f} \times h$$

Notes

where:

L = Lower boundary of the median class = 30

N = Total number of observations = 170

C = Cumulative frequency of the class before the median class = $30+f_1$.

f = Frequency of the median class = 40

h = Class width = 10

The median is given as 35, so:

$$35 = 30 + \frac{\left(\frac{170}{2}\right) - (30 + f_1)}{40} \times 10$$

Simplify this equation,

$35 = 30+(85-30-f_1)/4)$

$20 = 55-f_1$

$f_1 = 55-20=35$

To Find the Second Missing Frequency f_2, we know that the total frequency N = 170

$10 + 20 + f_1 + 40 + f_2 + 25 + 15 = 170$

Substitute $f_1=35$

$10 + 20 + 35 + 40 + f_2 + 25 + 15 = 170$

$145 + f_2 = 170$

$f_2 = 170 -145 = 25$

The missing frequencies are: $f_1 = 35$ and $f_2 = 25$.

Example 20: Calculate median from the following data:

Mid-value	15	25	35	45	55	65	75	85
Frequency	5	9	13	21	20	15	8	3

Solution: To calculate the median from the given data, follow these steps:

1. **Determine the Class Interval**

   The class interval (h) can be found by observing the difference between consecutive mid-values.

   Class Interval (h) = Difference between consecutive mid-values

   = 25−15 = 10

   So, the class interval is 10.

Notes

2. Calculate the Cumulative Frequency for each mid-value:

- Cumulative frequency up to 15: 5
- Cumulative frequency up to 25: 5+9=14
- Cumulative frequency up to 35: 14+13=27
- Cumulative frequency up to 45: 27+21=48
- Cumulative frequency up to 55: 48+20=68
- Cumulative frequency up to 65: 68+15=83
- Cumulative frequency up to 75: 83+8=91
- Cumulative frequency up to 85: 91+3=94

3. Determine the Median Class

The total number of observations (N) is 94. Median position = N/2 = 94/2 = 47.

The cumulative frequency just greater than 47 is 48, which corresponds to the mid-value of **45**. Since the class interval is 10, the class interval for the median class is: 40−50

4. Apply the Median Formula

The formula for the median in grouped data is:

$$Median = L + \frac{\left(\frac{N}{2}\right) - C}{f} \times h$$

where,

Lower boundary (L) = 40

Frequency of the median class (f) = 21

Cumulative frequency before the median class (C) = 27

Class interval (h) = 10

Substituting the values:

$$Median = 40 + \frac{47 - 27}{21} \times 10$$

$$= 40 + (20/21) \times 10$$

$$= 40 + (0.952) \times 10$$

$$= 40 + 9.52 \approx 49.52$$

Notes

Example 21: How are the mean and median affected when it is known that for a group of 10 students, scoring an average of 60 marks, the best paper was wrongly marked 80 instead of 75?

Solution: Here's how to assess the impact on the mean and median when correcting a score:

1. **Initial Data:**
   - Average (Mean) score = 60
   - Number of students = 10
   - Total score (Sum) = Average × Number of students = 60 × 10 = 600
2. **Error in Scoring:**
   - The best paper was wrongly marked 80 instead of 75.
3. **Correct Total Score:**
   - Correct total score = 600 - 80 + 75 = 595
4. **Correct Mean:** Correct Mean = Correct Total Score/Number of Students = 595/10 = 59.5

The median is the middle value of the dataset when it is ordered. For a dataset with 10 students, the median will be the average of the 5th and 6th values when ordered. If the incorrect score (80) is one of the highest values, replacing it with 75 might not affect the median if 75 falls among the lower half of the scores. However, if 75 is a lower value and does not change the middle position, the median remains the same.

Example 22: Find out the median of the following series:

Wages (in Rupees)	No. of Labourers
260-270	5
250-260	10
240-250	20
230-240	5
220-230	3

Notes

Solution:

To find the median of the given series, follow these steps:

1. **Calculate the Cumulative Frequencies:**
   - For the interval 220-230: Cumulative Frequency = 3
   - For the interval 230-240: Cumulative Frequency = 3 + 5 = 8
   - For the interval 240-250: Cumulative Frequency = 8 + 20 = 28
   - For the interval 250-260: Cumulative Frequency = 28 + 10 = 38
   - For the interval 260-270: Cumulative Frequency = 38 + 5 = 43

Wages (in Rupees) (m)	No. of Labourers (f)	Cumulative Frequency (c.f.)
260-270	5	5
250-260	10	15
240-250	20	35
230-240	5	40
220-230	3	43
	n = 43	

2. **Determine the Median Class:**
   - Total Number of Labourers = 43
   - Median position = 43/2 = 21.5
   - The median class is the class interval where the cumulative frequency just exceeds 21.5. Here, it is the interval 240 – 250 because the cumulative frequency is 28 (which is greater than 21.5).

3. **Use the Median Formula:**

$$Median = L + \frac{\left(\frac{N}{2}\right) - C}{f} \times h$$

where:

L = Lower boundary of the median class = 240

N = Total number of labourers = 43

C = Cumulative frequency of the class before the median class = 8

f = Frequency of the median class = 20

h = Class interval width = 10 (e.g., 250 – 240)

Notes

Substituting the values:

$$Median = 240 + \frac{(21.5) - 8}{20} \times 10$$

$$= 240 + (13.5/20) \times 10$$

$$= 240 + 6.75 = 246.75.$$

So, the median of the series is 246.75 Rupees.

Example 23: Calculate the mode from the following data:

Income	15-24	25-34	35-44	45-54	55-64	65-74
No. of Workers	8	10	15	25	40	20

Solution:

To calculate the mode for the given data, follow these steps:

1. **Identify the Modal Class:** The modal class is the class interval with the highest frequency. Here, the highest frequency is 40, which corresponds to the class interval **55-64**. So, the modal class is **55-64**.
2. **Apply the Mode Formula:** The formula for the mode in grouped data is:

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Where:

l = Lower boundary of the modal class = 55

f_1 = Frequency of the modal class = 40

f_0 = Frequency of the class before the modal class = 25

f_2 = Frequency of the class after the modal class = 20

h = Class interval width = 10 (since 25-34, 35-44, etc., all have a width of 10)

Step 3: Substitute the Values into the Formula

$$Mode = 55 + \frac{40 - 25}{2(40) - 25 - 20} \times h$$

$$= 55+(15/(80-45))\times 10$$

$$= 55+(15/35) \times 10$$

$$= 55+(0.4286) \times 10$$

$$= 55+4.29 \approx 59.29.$$

Notes

Example 24: Find the missing frequency for the following incomplete distribution by using the appropriate formula when mode is 36.

Variable	Frequency
0-10	5
10-20	7
20-30	**?**
30-40	?
40-50	10
50-60	6
	50

Solution: We are given that the mode is 36, and the total frequency is 50. Let the missing frequency corresponding to class 20-30 and 30-40 be x and y respectively. Let's follow the steps to calculate the missing frequencies.

1. **Identify the Modal Class:** The mode is given as 36, which falls in the class interval **30-40**. Therefore, the modal class is **30-40**.
2. **Use the Total Frequency:** We know that the total frequency is 50:

$$5+7+x+y+10+6=50$$

$$28+x+y=50$$

$$x+y=22$$

$$y=22-x$$

Step 3: Apply the Mode Formula

The formula for the mode in grouped data is:

$$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Where:

l = Lower boundary of the modal class = 30

f_1 = Frequency of the modal class (30-40) = y = 22-x

f_0 = Frequency of the class before the modal class (20-30) = x

f_2 = Frequency of the class after the modal class (40-50) = 10

h. = Class interval width = 10 (since 25-34, 35-44, etc., all have a width of 10)

Notes

Substitute the Values into the Formula

Given that the mode is 36:

$$36 = 30 + \frac{(22-x)-x}{2(22-x)-x-10} \times 10$$

$$6 = (22 - 2x/34 - 3x) \times 10$$

$$0.6 = (22 - 2x/34 - 3x)$$

$$0.6 \times (34 - 3x) = 22 - 2x$$

$$2x - 1.8x = 22 - 20.4$$

$$0.2x = 1.6$$

$$x = 1.6/0.2$$

$$x = 8$$

$$y = 22\text{-}x = 22\text{-}8 = 14$$

The missing frequencies are: x = 8 and y = 14.

Example 25: The average marks secured by 50 students was 44. Later on, it was discovered that a score 36 was misread as 56. Find the correct average secured for the students.

Solution:

Given N = 50 and mean or $\bar{X}$ = 44

$$\bar{X} = \frac{\Sigma X}{N} = \Sigma X = \bar{X}N \text{ or } 44 \times 50 = 2200$$

However, since a score of 36 was misread as 56, the total of $N\bar{X}$ should have been 2200 – 56 + 36 = 2180. Hence the corrected mean would be

$$\overline{X} = 2180/50 = 43.6 \text{ marks.}$$

4.8 Exercise

1. Define an Average. Describe briefly advantages and disadvantages of Arithmetic Mean.
2. What are the mathematical properties of Arithmetic Mean.
3. The heights (in cms) of 10 students of a class were noted as shown below. Compute the arithmetic mean.

Notes

S. No.	1	2	3	4	5	6	7	8	9	10
Height	160	167	174	158	155	171	162	152	156	175

4. Determine median from the following data:

30, 37, 54, 58, 61, 64, 31, 34, 52, 55, 32, 62, 28, 47, 55

5. Determine the mode of the following data:

58, 60, 31, 62, 48, 37, 78, 43, 65, 48

6. The following table shows the distribution of the number of students per teacher in 750 colleges

Students	1	4	7	10	13	16	19	22	25	28
Frequency	7	46	165	195	189	89	28	19	9	3

Calculate arithmetic mean, median and mode.

7. Calculate arithmetic mean, median and mode for the following data.

Marks Obtained	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of Students	8	14	22	26	15	10	5

8. Calculate the median, quartiles, 4^{th} decile and 60^{th} percentile the following data

Marks Obtained	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of Students	8	14	22	26	15	10	5

9. Calculate the median income for the following distribution:

Income (Rs.)	40 –45	45 – 50	50 – 55	55 – 60	60 – 65	65 – 70	70 - 75
No. of Persons	2	7	10	12	8	3	3

10. Find median, Q_1, Q_3, D_4, D_7, P_{26}, P_{45} and P_{70} from the following data

S. No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Marks:	5	12	17	23	28	31	37	41	42	49	54	58	65	68	17

11. The mean age of a group of 100 children was 9.35 years. The mean age of 25 of them was 8.75 years and that of another 65 was 10.51 years. What was the mean age for the remainder?

12. A firm of readymade garments makes both men's and women's shirts. Its profits average 6 percent of sales, its profits in men's shirts average 8 percent of sales; and women's shirts comprise 60 percent of output. What is the average profits per sales rupee in women's shirts?

Notes

13. The mean wage of 150 labourers working in a factory running three shifts of 60, 40, and 50 labourers is Rs. 114.00. The mean wage of 60 labourers working in the first shift is Rs. 121.50 and that of 40 labourers working in the second shift is Rs. 107.75. Find the mean wage of the labourers working in the third shift.

14. A market with 168 operating firms has the following distribution of average number of workers in various income groups:

Income Groups	150-300	300-500	500-800	800-1200	1200-1800
No. of Firms	40	32	26	28	42
Average No. of Workers	8	12	7.5	8.5	4

Find the average salary paid in the whole market.

15. The expenditure of 1000 families is given as under:

Expenditure (in Rs.)	40—59	60—79	80—99	100—119	120—139
No. of Families	50	?	500	?	50

The median and mean for the distribution are both Rs. 87.50 respectively. Calculate the missing frequencies.

16. The median and mode of the following wage distribution are known to be Rs. 335 and Rs. 340 respectively. Three frequency values from the table, however, are missing. Find the missing values.

Wages in Rs.	**Frequency**
0-100	4
100-200	16
200-300	60
300-400	?
400-500	?
500-600	?
600-700	4
	230

17. From the following data related to unemployment, calculate the standardized unemployment rate, the standardized rate of unemployment of local population and the crude rate of unemployment of local population.

Notes

Age (Years)	Standard Population		Local Population	
	Unemployment	Percentage	Unemployment	Percentage
16-30	250	5%	300	4%
30-45	350	8%	300	9%
45-60	300	12%	350	12%
60 and Over	100	15%	50	20%
Total	1000		1000	

18. Calculate the value of mode by the usual formula (after grouping if necessary):

Class - Interval	Frequency
10-20	4
20-30	6
30-40	5
40-50	10
50-60	20
60-70	22
70-80	24
80-90	6
90-100	8
100-110	1

19. A distribution consists of three components with total frequencies of 200, 250 and 300 having means 25, 10 and 15 respectively. Find the mean of combined distribution.

20. In a moderately skewed distribution:

(a) Arithmetic mean = 24.6 and the mode = 26.1. Find the value of the median and explain the reason for the method employed.

(b) In a moderately asymmetrical distribution, the value of the median is 42.8 and the value of the mode is 40. Find the mean.

(c) If the mode and mean of a moderately asymmetrical series are respectively 26 and 20.2 meters, compute the most probable median.

21. Graphically with the help of the ogive curve, find the value of Q_1, Q_3, median, P_{40} and D_6

Class Interval	10-14	15-19	20-24	25-29	30-34	35-39
Frequency	5	10	15	20	10	5

LESSON

5

Measures of Dispersion

STRUCTURE

5.1 Dispersion

Averages discussed earlier fail to reveal the full details of the distribution. Two or three distributions may have the same average but still they may differ from each other in many ways.

Suppose, there are three series of nine items each as follows:

Series A	Series B	Series C
50	48	5
50	50	15
50	46	20
50	49	25
50	47	35
50	52	80
50	53	85
50	51	90
50	54	95

In the series A, the mean is 50 and the value of all the items is identical. The items are not at all scattered, and the mean is the representative of this distribution. However, in the series B, though the mean is 50 yet all the items of the series have different values.

Notes

But the items are not very much scattered as the minimum value of the series B is 46 and the maximum is 54 in the range. In the series C also, the mean is 50 and the values of different items are also different, but here the values are very widely scattered. Though the mean is the same in all the three series, yet the series differ widely from each other in their formation. Obviously, the average is not giving us the complete information about the series. In such cases, further Statistical analysis of the data is necessary so that these differences between various series can be studied and accounted for. Such analysis will make our results more accurate and we shall be more confident of our conclusions. As we can see the spread among the items in the Series A is zero, Series B varies within a small range, while in the Series C the values are widely scattered. It is evident from the above, that a study of the extent of the scatter around average should also be made to throw more light on the composition of a series.

This spread or scatteredness of the data is called dispersion.

Some important definitions of dispersion are given below:

"Dispersion or spread is the degree of the scatter or variation of the variable about a central value."

– Brooks and Dick

"Dispersion is the measure of the variations of the items."

– A. L. Bowley

"The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data."

– Spiegel

"Measures of variability are usually used to indicate how tightly bunched the sample values are around the mean."

– Dyckman and Thomas

From the above definitions, it is clear that in a general sense the term dispersion refers to the variability in the size of the items. If the variation is substantial, dispersion is said to be significant and if the variation is very little, dispersion is said to be insignificant.

Notes

Measures of Dispersion

Dispersion means scatteredness. If gives us an idea about the homogeneity or heterogeneity of the distribution. Measure of Dispersion indicates the disparity of data from one another. It is required to provide accurate view of their distribution.

5.1.1 Characteristics for an Ideal Measure of Dispersion

Same as that for ideal measure of central Tendency

1. Rigidly defined
2. Easy to calculate and understand
3. Based on all the observations
4. Suitable for further mathematical treatment
5. Should be least affected by fluctuations of sampling

5.1.2 Measures of Dispersion – are Following

1. Range
2. Quartile deviation or semi-interquartile range
3. Mean deviation
4. Standard deviation

5.2 Range

It is the difference between the greatest and the smallest observations of the distribution.

If L is maximum value, S – minimum value

Then, Range = L – S.

It is the simplest but a crude measure of dispersion. It is based on two extreme observations which are subject to chance fluctuations, hence it is not at all a reliable measure of dispersion. This is an absolute measure of dispersion and is not suitable for comparison in case distributions are in different units. For comparison, a relative measure is used, called coefficient of range.

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

Notes

5.2.1 Merits and Demerits of Range

Merits:

1. It is the simplest method of studying variation.
2. It is easy to compute.
3. It is a quick method and takes less time for computation.

Demerits:

1. It is subjected to fluctuations of sampling.
2. It is affected by extreme observations.
3. It is not based on each and every item of the distribution.
4. It cannot be computed in open-end distributions.

Example 1: Find the range of given observations: 32, 41, 28, 54, 35, 26, 23, 33, 38, 40.

Solution:

Since 23 is the lowest value and 54 is the highest value, therefore, the range of the observations will be;

Range (X) = Max (X) – Min (X)

= 54 – 23

= 31

Example 2: Following are the marks of students in Mathematics: 50, 53, 50, 51, 48, 93, 90, 92, 91, 90. Find the range of the marks. Also obtain coefficient of range

Solution:

The range of marks will be:

Range = Maximum marks – Minimum marks

Range = 93 – 48 = 45

$$\text{Coefficient of Range} = \frac{93-48}{93+48} = \frac{45}{141} = 0.319$$

5.3 Interquartile Range and Quartile Deviation

Suppose Q_1 is the first quartile, Q_2 is the median or second quartile, and Q_3 is the third quartile for the given data set then

Notes

Interquartile Range = $Q_3 - Q_1$

The Quartile Deviation can be defined mathematically as half of the difference between the first and the third quartile. Here, quartile deviation can be represented as QD.

Quartile Deviation is also known as the Semi Interquartile range.

$QD = (Q_3 - Q_1)/2$

Quartile Deviation gives the average amount by which the two quartiles differ from the median. It is an absolute measure of dispersion. The relative measure of dispersion is the coefficient of Quartile Deviation given as:

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

5.3.1 Merits and Demerits of Quartile Deviation

Merits:

1. It is used in measuring the variation in case of open-end distributions as it is a positional measure of dispersion.
2. It is not affected by extreme observations and is therefore used in badly skewed distributions. It is useful in badly skewed distributions.

Demerits:

1. It does not depend on each and every item of the distribution as it ignores the first 25% and last 25% of items.
2. It is not capable of further algebraic treatment.
3. It is affected by fluctuations in sampling.

Example 3: Find the quartiles and quartile deviation of the following data:

17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Solution:

Given data:

17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Ascending order of the given data is:

2, 5, 7, 7, 8, 8, 10, 10, 14, 15, 17, 18, 24, 27, 28, 48

Number of data values = n = 16

Q_2 = Median of the given data set

Notes

n is even, median = (1/2) [$(n/2)^{th}$ observation and $(n/2 + 1)^{th}$ observation]

= (1/2) [8th observation + 9th observation]

= (10 + 14)/2

= 24/2

= 12

$Q_2 = 12$

Now, lower half of the data is:

2, 5, 7, 7, 8, 8, 10, 10 (even number of observations)

Q_1 = Median of lower half of the data

= (1/2) [4th observation + 5th observation]

= (7 + 8)/2

= 15/2

= 7.5

Also, the upper half of the data is:

14, 15, 17, 18, 24, 27, 28, 48 (even number of observations)

Q_3 = Median of upper half of the data

= (1/2) [4th observation + 5th observation]

= (18 + 24)/2

= 42/2

= 21

Quartile deviation = $(Q_3 - Q_1)/2$

= (21 – 7.5)/2

= 13.5/2

= 6.75

Therefore, the quartile deviation for the given data set is 6.75.

5.4 Mean Deviation

The mean deviation is the average difference between the items in a distribution from the mean of that series. Hence it is also known as average deviation.

Notes

Mean deviation is least when taken from median. However, in practice, arithmetic mean is most frequently used in calculating mean deviation.

It is a better measure of dispersion as it is based on all the observations. But since it ignores the signs of the deviations it becomes useless for further mathematical treatment.

Mean deviation in case of individual observations:

If x_i, i = 1, 2, ... n are the observations then mean deviation from the average A is

$$\text{Mean deviation (M.D.)} = \frac{\sum|x_i - A|}{N} = \frac{\sum|D|}{N}$$

where $| x_i - A | = |D|$ is the absolute value of $(x_i - A)$ (ignoring the sign) 'A' is usually taken as mean, median or mode.

Mean deviation in case of discrete or continuous distribution:

If x_i/f_i ; I = 1, 2...., n is the frequency distribution then mean deviation from the average A is

$$\text{M. D.} = \frac{\sum f|x_i - A|}{\sum f} = \frac{\sum f|D|}{N}$$

Coefficient of mean deviation:

It is the relative measure of mean deviation

$$\text{Coefficient of M. D.} = \frac{M.D.}{Average\ from\ which\ mean\ deviation\ is\ calculated}$$

5.4.1 Merits and Demerits of Mean Deviation

Merits:

1. It is simple to understand
2. It is easy to compute
3. It is based on all the observations of the series
4. It is not affected much by the extreme observations
5. It is used to forecast business cycle by national Bureau of Economic Research

Notes

Demerits:

1. It is not capable of further algebraic treatments since it ignores sign
2. It does not always give accurate results. It gives the best result when the deviations are taken from the median
3. It is rarely used in sociological studies. Hence it has limited use.

Example 4: Determine the mean deviation about mean for the data values 5, 3,7, 8, 4, 9.

Solution:

Given data values are 5, 3, 7, 8, 4, 9.

First, find the mean for the given data:

Mean, $\mu = (5+3+7+8+4+9)/6$

$\mu = 36/6$

$\mu = 6$

Therefore, the mean value is 6.

Now, subtract the mean from each of the data value, and ignore the minus symbol if any

The obtained values are 1, 3, 1, 2, 2, 3.

Therefore, the mean deviation is

$= (1 + 3 + 1 + 2 + 2 + 3)/6$

$= 12/6$

$= 2$

5.5 Standard Deviation and Root Mean Square Deviation

Standard Deviation (S.D.) measures the absolute dispersion or variability of a distribution. The greater the variability the greater the value of standard deviation. Smaller the value of standard deviation, higher is the degree of uniformity and homogeneity among the observations. It satisfies most of the properties of a good measure of dispersion hence it is by far most important and widely used measure.

Standard deviation is the positive square root of the arithmetic mean of the squares of the deviation of the given values from their arithmetic mean hence it is also called **root mean square deviation**. It is denoted by σ.

Notes

5.5.1 Standard Deviation in Case of Individual Observations is obtained using any of the following formula:

(i) $\sigma = \sqrt{\frac{1}{N}\sum(x-\bar{x})^2}$ *or* $\sqrt{\frac{\sum x^2}{N} - \bar{x}^2}$

(ii) $\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$

where, d = X – A, A is any assumed mean, N is total number of observations.

5.5.2 Standard Deviation in Case of Discrete Series is obtained using any of the following formula:

***(i)* Actual mean method**

$$\sigma = \sqrt{\frac{1}{N}\sum f(x-\bar{x})^2} \quad \text{or} \quad \sqrt{\frac{\sum fx^2}{N} - \bar{x}^2}$$

Where N = $\sum f$

***(ii)* Assumed mean method**

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Where, d = x – A, A is assumed mean and N = $\sum f$

***(iii)* Step deviation method**

$$\sigma = h \times \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2}$$

Where, $u = \frac{x-A}{h}$, A is assumed mean and h is class size

5.5.3 Standard Deviation in Case of Continuous Series is obtained using any of the following formula:

For the frequency distribution x_i/f_i ; i = 1, 2, ..., n

$$\sigma = h \times \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2}$$

Notes

Where, $u = \frac{m-A}{h}$, A is assumed mean, m is mid-value of the class and h is class size

5.5.4 Merits and Demerits of Standard Deviation:

Merits

(i) The step of squaring the deviations $(x_i - \bar{x})$ overcomes the drawback of ignoring the signs in mean deviation. Hence it is suitable for further mathematical of sampling.

(ii) Standard deviation is least affected by fluctuations of sampling.

(iii) It is based on all the observations of the data.

Demerits

(i) It is not readily comprehensible for a non-mathematical person since it requires finding of square root.

(ii) It gives more weights to extreme values.

5.5.5 Variance

The square of standard deviation is called the **variance.** It is denoted by σ^2

$$\sigma^2 = \frac{1}{N}\sum f\left(x-\bar{x}\right)^2 = \frac{1}{N}\sum fx^2 - \bar{x}^2$$

5.5.6 Coefficient of Variation

It is the relative measure of dispersion and is used in problems where we want to compare the variability of two or more series. The series with greater coefficient of variation is said to be more variable or inconsistent and the series with less coefficient of variation is said to be more homogeneous or consistent or more uniform or more stable. It is denoted by C.V. and given by the formula

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

Example 5: Calculate the mean, variance and standard deviation for the following data:

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	27	10	7	5	4	2

Notes

Solution:

Class Interval	Frequency (f)	Mid Value (x)	fx	fx^2
0 – 10	27	5	135	675
10 – 20	10	15	150	2250
20 – 30	7	25	175	4375
30 – 40	5	35	175	6125
40 – 50	4	45	180	8100
50 – 60	2	55	110	6050
	$\sum f = 55$		$\sum fx = 925$	$\sum fx^2 = 27575$

$N = \sum f = 55$

Mean $= (\sum fx_i)/N = 925/55 = 16.818$

Variance $= \frac{1}{N}\sum x_i^2 - \bar{x}^2$

$= 1/(55)\ [27575] - [16.818]^2$

$= 501.3636 - 282.8451$

$= 218.5185$

Standard deviation $= \sqrt{variance} = \sqrt{218.5185} = 14.782$

Example 6: Compute the standard deviation for the following frequency distribution:

Class interval	0–4	4–8	8–12	12–16
Frequency	4	8	2	1

Solution:

Class Interval	f	x	fx	fx^2
0–4	4	2	8	16
4–8	8	6	48	288
8–12	2	10	20	200
12–16	1	14	14	196
	f = 15		fx = 90	$fx^2 = 700$

Variance $= 700/15 - (90/15)^2 = 10.66$

Standard Deviation $= \sqrt{variance} = \sqrt{10.66} = 3.26$

Notes

5.6 Miscellaneous Examples

Example 7: Determine range and coefficient of range from the following data:

Wages	100-110	110-120	120-130	130-140	140-150	150-160
No. of Workers	10	12	15	16	9	8

Solution:

Minimum value = 100

Maximum value = 160

Range = L – S = 160 – 100

Coefficient of range = $\frac{L-S}{L+S}=\frac{160-100}{160+100}=\frac{60}{260}=0.23$

Example 8: Find the interquartile range and coefficient of quartile deviation from the data given below:

200, 210, 208, 160, 220, 250, 300

Solution: Arranging the data in ascending order, we get

160, 200, 208, 210, 220, 250, 300

N = 7

Q_1 = size of $(\frac{N+1}{4})^{th}$ item

$= \frac{7+1}{4} = 2^{nd}$ item = 200

Q_3 = size of $\left(\frac{3(N+1)}{4}\right)^{th}$ item

$= 6^{th}$ item = 250

Interquartile range = $Q_3 - Q_1$

= 250 – 200

= 50

Notes

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{250 - 200}{250 + 200}$$

$$= \frac{50}{450}$$

$$= 0.11$$

Example 9: Find the interquartile range from the following data:

Marks (more than)	0	20	40	60	80	100	120
No. of Students	80	76	50	28	18	9	3

Solution:

Marks	Frequency (f)	Cumulative Frequency (c.f.)
0 – 20	4	4
20 – 40	26	30
40 – 60	22	52
60 – 80	10	62
80 – 100	9	71
100 – 120	6	77
120 – 140	3	80
	N = 80	

N = 80

Q_1 = size of $(\frac{N}{4})^{th}$ item

= $\frac{80}{4}$ = 20^{nd} item

Therefore, the quartile class is 20 - 40

$$Q_1 = l + \frac{\frac{N}{4} - cf}{f} \times h = 20 + \frac{20 - 4}{26} \times 20 = 20 + 12.3 = 32.3$$

Q_3 = size of $\left(\frac{3N}{4}\right)^{th}$ item

= 60^{th} item

Notes

Therefore, quartile class is 60 – 80

$$Q_3 = l + \frac{\frac{3N}{4} - cf}{f} \times h = 60 + \frac{60-52}{10} \times 20$$

$$= 60 + 16 = 76$$

Interquartile range = $Q_3 - Q_1$

= 76 – 32.3

= 43.7

Example 10: Calculate the coefficient of mean deviation from the data given below:

Mid value	115	125	135	145	155	165	175	185	195
Frequency	6	25	48	72	116	60	38	22	3

Solution:

Since we are given the mid-values, we first find the class intervals. Since the differences between the two consecutive mid-values is 10, therefore

Class size = 10

The first class-interval will be 110 -120. Subsequent class intervals can be found easily.

Class Intervals	**Mid-Value (x)**	**Frequency (f)**	**Cumulative Frequency (c.f.)**	**|d| = |x – 53.8|**	**f|d|**
110 – 120	115	6	6	38.8	232.8
120 – 130	125	25	31	28.8	720.0
130 – 140	135	48	79	18.8	902.4
140 – 150	145	72	151	8.8	633.6
150 – 160	155	116	267	1.2	139.2
160 – 170	165	60	327	11.2	672.0
170 – 180	175	38	365	21.2	805.6
180 – 190	185	22	387	31.2	686.4
190 - 200	195	3	390	41.2	123.6
		N = 390			$\Sigma f\|d\| = 4915.6$

Median = $\left(\frac{N}{2}\right)^{th}$ item = 195th item

Notes

Therefore, median class = 150 – 160

$$\text{Median} = l + \frac{\frac{N}{2} - cf}{f} \times h = 150 + \frac{195 - 151}{116} \times 10 = 150 + 3.79 = 153.8$$

$$\text{Mean deviation} = \frac{\Sigma f|d|}{N} = \frac{4915.6}{390} = 12.6$$

$$\text{Coefficient of mean deviation} = \frac{mean\,deviation}{median} = \frac{12.6}{153.8} = 0.0819$$

Example 11: Calculate the mean deviation from mean from the following data:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	4	6	10	20	10	6	4

Solution:

Marks	**Mid Value (m)**	**f**	**d = (m – 35)/ 10**	**fd**	**D = |d – mean|/10**	**f|D|**
0 – 10	5	4	-3	-12	3	12
10 – 20	15	6	-2	-12	2	12
20 – 30	25	10	-1	-10	1	10
30 – 40	35	20	0	0	0	0
40 – 50	45	10	1	10	1	10
50 – 60	55	6	2	12	2	12
60 - 70	65	4	3	12	3	12
		N = 60		Σfd = 0		Σf|D| = 68

$$\text{Mean} = A + \frac{\sum fd}{N} \times i$$

$$= 35 + \frac{0}{60} \times 10 = 35$$

$$\text{Mean deviation} = \frac{\sum f|D|}{N} \times i = \frac{68}{60} \times 10 = 11.33$$

Example 12: Find the standard deviation for the following distribution:

X	4.5	14.5	24.5	34.5	44.5	54.5	64.5
f	1	5	12	22	17	9	4

Notes

Solution:

X	f	d = (X – 34.5)/10	fd	fd^2
4.5	1	-3	-3	9
14.5	5	-2	-10	20
24.5	12	-1	-12	12
34.5	22	0	0	0
44.5	17	1	17	17
54.5	9	2	18	36
64.5	4	3	12	36
	N = 70		Σfd = 22	$\Sigma fd^2 = 130$

$$\sigma = i \times \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

$$= 10 \times \sqrt{\frac{130}{70} - \left(\frac{22}{70}\right)^2}$$

$$= 10 \times \sqrt{1.757} = 13.26$$

Example 13: The number of employee, wages per employee and the variance of the wages of employees for two factories are given below:

	Factory A	**Factory B**
No. of Employees	50	100
Average Wages per Employee per Month	1200	850
Variance of the Wages (Rs.)	81	256

(i) In which factory is the greater variation in the distribution of wages per employee?

(ii) Suppose I factory B, the wages of an employee were wrongly noted as Rs. 900 instead of Rs. 910. What would be the correct variance of wages in Factory B?

Solution:

(i) In order to compare the distribution of wages in the two factories, we find coefficient of variation. The Coefficient of variation is given by the formula

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

Notes

Now,

C.V. for factory A = $\frac{9}{1200}\times 100 = 0.75\%$

C.V. for factory B = $\frac{16}{850}\times 100 = 1.88\%$

Since the coefficient of variation is greater for factory B, it shows greater variation in the distribution of wages per employee.

(ii) For factory B,

$$\Sigma X = 850\times 100 = 85000$$

Corrected $\Sigma X = 85000 - 900 + 910 = 85010$

Thus the corrected mean, $\bar{X} = \frac{85010}{100} = 850.10$

Now,

Variance = $\sigma^2 = \frac{\sum X^2}{N} - \left(\bar{X}\right)^2$

It is given for factory B, variance = 256, N = 100, mean = 850

Hence, from the variance formula, we get

$$256 = \frac{\sum X^2}{100} - \left(850\right)^2$$

25600 = $\sum X^2$ – 72250000

$\sum X^2$ = 72250000 – 25600 = 72275600

Hence, correct $\sum X^2$ = 72275600 – $(900)^2$ + $(910)^2$

= 72293700

Now the correct variance is

$$correct\ variance = \frac{correct \sum X^2}{N} - \left(correct\ \bar{X}\right)^2$$

$$= \frac{72293700}{100} - \left(850.10\right)^2$$

$$= 266.99$$

Notes

Example 14: Calculate the variance of the following data:

Variable	20 - 25	25 - 30	30 - 35	35 - 40	40 - 45	45 - 50
Frequency	170	110	80	45	40	35

Solution:

Variable	**Mid-point (m)**	**Frequency (f)**	**u = $\frac{m-32.5}{5}$**	**fu**	**fu^2**
20 – 25	22.5	170	-2	-340	680
25 – 30	27.5	110	-1	-110	110
30 – 35	32.5	80	0	0	0
35 – 40	37.5	45	1	45	45
40 – 45	42.5	40	2	80	160
45 - 50	47.5	35	3	105	315
		N = 480		$\Sigma fu = -220$	$\Sigma fu^2 = 1310$

Variance = $\sigma^2 = \left[\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2\right] \times h^2$

$= \left[\frac{1310}{480} - \left(\frac{-220}{480}\right)^2\right] \times 5^2$

$= (2.729 - 0.21) \times 25 = 62.975$

Example 15: From the data given below state which series is more consistent:

Variable	**Series A**	**Series B**
10 – 20	10	18
20 – 30	18	22
30 – 40	32	40
40 – 50	40	32
50 – 60	22	18
60 – 70	18	10

Notes

Solution:

Variable	Mid Value (x)	Series A			Series B		
		f	fx	fx^2	f	fx	fx^2
10 – 20	15	10	150	2250	18	270	4050
20 – 30	25	18	450	11250	22	550	13750
30 – 40	35	32	1120	39200	40	1400	49000
40 – 50	45	40	1800	81000	32	1440	64800
50 – 60	55	22	1210	66550	18	990	54450
60 – 70	65	18	1170	76050	10	650	42250
		N = 140	Σfx = 5900	Σfx^2 = 276300	140	Σfx =5300	Σfx^2 =228300

Mean for series A = $\frac{\Sigma fx}{N} = \frac{5900}{140} = 42.14$

Standard deviation for series A = $\sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2}$

$= \sqrt{\frac{276300}{140} - \left(\frac{5900}{140}\right)^2} = 14.05$

C.V. of series A = $\frac{\sigma}{\bar{x}} \times 100 = \frac{14.05}{42.14} \times 100 = 33.34\%$

Now,

Mean for series B = $\frac{\Sigma fx}{N} = \frac{5300}{140} = 37.86$

Standard deviation for series B = $\sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2}$

$= \sqrt{\frac{228300}{140} - \left(\frac{5300}{140}\right)^2} = 14.06$

C.V. of series B = $\frac{\sigma}{\bar{x}} \times 100 = \frac{14.06}{37.86} \times 100 = 37.14\%$

Since the coefficient of variation is less for series A hence series A is more consistent.

Notes

5.7 Exercise

1. What is Dispersion? Discuss the Merits and Demerits of (i) Range (ii) Mean Deviation.
2. Find the range for the following data:
   (a) 63, 89, 98, 125, 79, 108, 117, 68
   (b) 43.5, 13.6, 18.9, 38.4, 61.4, 29.8
3. A teacher asked the students to complete 60 pages of a record note book. Eight students have completed only 32, 35, 37, 30, 33, 36, 35 and 37 pages. Find the standard deviation of the pages yet to be completed by them.

Mass in kg.	60–62	63–65	66–68	69–71	72–74
Number of Students	5	18	42	27	8

4. From the following frequency distribution, compute the standard deviation of 100 students:
5. Calculate the mean and standard deviation for the following data:

Size of Item	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

6. What is the range for the following data set:
   1,2,8,9,7,4,1,1,3,2,3
7. What is the range of the data sets 6, 2, 11, 14, 19, and 15?
8. Determine the highest value in the data set, if the range equals 40 and the lowest value should be equal to 6.
9. Find the range of the first 5 composite numbers.
10. Determine the interquartile range value for the first ten prime numbers.
11. Find the variance for an ungrouped data 5,12,3,18,6,8,2,10.
12. Find the variance of the following distribution.

Class Interval	Frequency
20 - 24	15
25 - 29	25
30 - 34	28
35 - 39	12
40 - 44	12
45 - 49	8

Notes

13. During the 10 weeks of a session, the marks obtained by two candidates, Ramesh and Suresh, taking the computer programme course are given below:

Ramesh	58	59	60	54	65	66	52	75	69	52
Suresh	87	89	78	71	73	84	65	66	56	46

(i) Who is the better scorer – Ramesh or Suresh?

(ii) Who is more consistent?

14. From the prices of shares of X and Y given below, state which share is more stable in value:

X	55	54	52	53	56	58	52	50	51	49
Y	108	107	105	105	106	107	104	103	104	101

15. Calculate the mean deviation from the median for the following data:

Age (yrs.)	4-6	6-8	8-10	10-12	12-14	14-16	16-18
No. of Students	30	90	120	150	80	60	20

16. A factory produces two types of electric lamps A and B. in an experiment relating to their life, the following results were obtained:

Length of Life (in hrs)	**No. of Lamps A**	**No. of Lamps B**
500 – 700	5	4
700 – 900	11	30
900 – 1100	26	12
1100 – 1300	10	8
1300 - 1500	8	6

Compare the variability of the life of the two varieties using coefficient of variation.

17. Find out who is better and consistent Batsman from the following data:

Batsman A	10	12	80	70	60	100	0	4
Batsman B	8	9	7	10	5	9	10	8

18. The mean and standard deviation of 15 items were found to be 8 and 2 respectively. On checking, it was discovered that one item 11 has been misread as 5. Calculate the correct mean and standard deviation.

LESSON

6

Moments

STRUCTURE

6.1 Moments

In statistics, "moments" are quantitative measures related to the shape of a set of points. Moments are used in various fields such as physics, engineering, and probability theory to understand the distribution and characteristics of data sets. First four moments are of key importance in statistics. These moments provide a comprehensive summary of the data's characteristics.

The first moment about the origin is the mean, which is a measure of central tendency.

The second moment about the mean is the variance, which measures the spread or dispersion of the data.

The third moment about the mean is skewness, which measures the asymmetry of the data distribution.

The fourth moment about the mean is kurtosis, which measures the "tailedness" or peakedness of the data distribution. That is, it shows the presence of outliers and the shape of the data distribution's tails.

6.1.1 Moments About Mean

The rth moment of a variable x about the mean $\overline{x}$ is denoted by μ_r. It is given by the formula

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - \overline{x}\right)^r$$

Notes

Or

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (z_i)^r ,$$

where, $z_i = x_i - \bar{x}$

Here, f_i is the frequency of the ith class and N is the total frequency.

Keeping r = 1, 2, 3, 4 gives the first four moments about mean.

In particular, for r = 0

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^0 \Rightarrow \mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i = 1$$

For r = 1

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^1 = 0$$

For r = 2

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^2 = \sigma^2 = variance$$

For r = 3

$$\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^3$$

For r = 4

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^4$$

6.1.2 Moments about a point or Raw moments

The r[th] moment of a variable x about any point x = A, is denoted by μ'_r and is given by the formula,

$$\mu'_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^r ,$$

where, N = Σf_i

Or

Notes

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i \left(d_i\right)^r ,$$

where, $d_i = x_i - A$

Keeping r = 1, 2, 3, 4 in the above formula, we get the first four raw moments about any point A.

Thus,

For r = 1

$$\mu_1' = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - A\right)^1$$

For r = 2

$$\mu_2' = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - A\right)^2$$

For r = 3

$$\mu_3' = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - A\right)^3$$

For r = 4

$$\mu_4' = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - A\right)^4$$

Note: $\bar{x} = A + \frac{1}{N}\sum_{i=1}^{n} f_i d_i = A + \mu_1'$, where $d_i = x_i - A$

6.1.3 Moments About Origin

The first four moment about origin is obtained on keeping A = 0 in the formula for moment about a point.

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - A\right)^r = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - 0\right)^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i\right)^r$$

Notes

Thus, for r = 1

$$\mu_1' = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i\right)^1 = \overline{x} = mean$$

For r = 2

$$\mu_2' = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i\right)^2$$

For r = 3 and r = 4, we get respectively

$$\mu_3' = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i\right)^3, \qquad \mu_4' = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i\right)^4$$

6.1.4 Relation between μ_r and μ_r'

The moments about mean can be expressed in terms of moments about any point, i.e., in terms of raw moments. The relationship is given as

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

Conversely, the moments about a point can be expressed as moments about mean by the following relationships

$$\mu_2' = \mu_2 + \mu_1'^2$$

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3$$

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^2 + \mu_1'^4$$

Example 1: Find the first four moments for the following individual series:

X	1	3	9	12	20

Solution:

Sl. No.	x	$x - \overline{x}$	$(x - \overline{x})^2$	$(x - \overline{x})^3$	$(x - \overline{x})^4$
1	1	– 8	64	– 512	4096
2	3	– 6	36	– 216	1296
3	9	0	0	0	0

4	12	3	9	27	81
5	20	11	121	131	14641
$n = 5$	$\Sigma x = 45$	$\Sigma (x - \bar{x}) = 0$	$\Sigma (x - \bar{x})^2 = 230$	$\Sigma (x - \bar{x})^3 = 630$	$\Sigma (x - \bar{x})^4 = 20114$

$$\bar{x} = \frac{\Sigma x}{n} = 9$$

$$\mu_1 = \frac{\sum_{i=1}^{5} (x_i - \bar{x})^1}{n} = 0$$

$$\mu_2 = \frac{\sum_{i=1}^{5} (x_i - \bar{x})^2}{n} = \frac{230}{5} = 46$$

$$\mu_3 = \frac{\sum_{i=1}^{5} (x_i - \bar{x})^3}{n} = \frac{630}{5} = 126$$

$$\mu_4 = \frac{\sum_{i=1}^{5} f_i (x_i - \bar{x})^4}{n} = \frac{20114}{5} = 4022.8$$

Example 2: Calculate the variance and third central moment from the following data:

x_i	0	1	2	3	4	5	6	7	8
f_i	1	9	26	59	72	52	29	7	1

Solution:

x_i	f_i	$x_i - 4$	$f_i (x_i - 4)$	$f_i (x_i - 4)^2$	$f_i (x_i - 4)^3$
0	1	– 4	– 4	16	– 64
1	9	– 3	– 27	81	– 243
2	26	– 2	– 52	104	– 208
3	59	– 1	– 59	59	– 59
4	72	0	0	0	0
5	52	1	52	52	52
6	29	2	58	116	232
7	7	3	21	63	189
8	1	4	4	16	64
	$\Sigma f_i = 256$		$\Sigma f_i (x_i - 4) = – 7$	$\Sigma f_i (x_i - 4)^2 = 507$	$\Sigma f_i (x_i - 4)^3 = – 37$

$$\mu_1' = \frac{1}{N}\sum_{i=1}^{8} f_i\left(x_i - 4\right)^1 = \frac{-7}{256}$$

$$\mu_2' = \frac{1}{N}\sum_{i=1}^{8} f_i\left(x_i - 4\right)^2 = \frac{507}{256}$$

$$\mu_3' = \frac{1}{N}\sum_{i=1}^{8} f_i\left(x_i - 4\right)^3 = \frac{-37}{256}$$

$$\mu_2 = \mu_2' - \mu_1'^2 = \frac{507}{256} - \left(\frac{-7}{256}\right)^2 = 1.98047 - 0.00075 = 1.97972$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \frac{-37}{256} - 3\left(\frac{507}{256}\right)\left(\frac{-7}{256}\right) + 2\left(\frac{-7}{256}\right)^3$$

$$= -0.14453 + 0.16246 - 0.00004$$

$$= 0.01789$$

Example 3: Calculate μ_1, μ_2, μ_3, μ_4 for the following frequency distribution:

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students	1	6	10	15	11	7

Solution: n = 6

Marks	**No. of Students** f	**Mid Value** x	fx	$x - \bar{x}$	$f(x - \bar{x})$	$f(x - \bar{x})^2$	$f(x - \bar{x})^3$	$f(x - \bar{x})^4$
0-10	1	5	5	–30	– 30	900	– 27000	810000
10-20	6	15	90	–20	– 120	2400	– 48000	960000
20-30	10	25	250	–10	– 100	1000	– 10000	100000
30-40	15	35	525	0	0	0	0	0
40-50	11	45	495	10	110	1100	11000	11000
50-60	7	55	385	20	140	2800	56000	112000
	N = Σ f = 50		Σ fx = 1750		Σ $f(x - \bar{x})$ = 0	Σ $f(x - \bar{x})^2$ = 8200	Σ $f(x - \bar{x})^3$= –18000	Σ $f(x - \bar{x})^4$ = 3100000

$$\bar{x} = \frac{\Sigma x}{n} = 9$$

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^1 = \frac{0}{50} = 0$$

Notes

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^2 = \frac{8200}{50} = 164$$

$$\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^3 = \frac{-18000}{50} = -360$$

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^4 = \frac{3100000}{50} = 62000$$

Example 4: The first three moments of a distribution, about the value '2' of the variable are 1, 16 and –40. Show that the mean is 3, variance is 15 and $\mu_3 = -86$.

Solution: We have

A = 2, $\mu'_1 = 1$, $\mu'_2 = 16$ and $\mu'_3 = -40$

$$\mu'_1 = \bar{x} - A \Rightarrow \bar{x} = \mu'_1 + A$$

$$= 1 + 2 = 3$$

we know that

Variance = $\mu'_2 - \mu'^2_1 = 16 - (1)^2 = 15$

$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1 = -40 - 3(16)(1) + 2(1)^3 = -40 - 48 + 2$

$= -86.$

Example 5: The first four moments of a distribution, about the value '35' are –1.8, 240, –1020 and 144000. Find the values of μ_1, μ_2, μ_3, μ_4.

Solution: We have, A = 35,

$\mu'_1 = -1.8$, $\mu'_2 = 240$, $\mu'_3 = -1020$ and $\mu'_4 = 144000$

$\mu_1 = 0$

$\mu_2 = \mu'_2 - \mu'^2_1 = 240 - (-1.8)^2 = 236.76$

$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1 = -1020 - 3(240)(-1.8) + 2(-1.8)^3 = 264.36$

$\mu_4 = \mu'_4 - 4\mu_3\mu'_1 + 6\mu_2\mu'^2_1 - 3\mu'^4_1$

$= 144000 - 4(-1020)(-1.8) + 6(240)(-1.8)^2 - 3(-1.8)^4 = 141290.11.$

Notes

6.2 Skewness

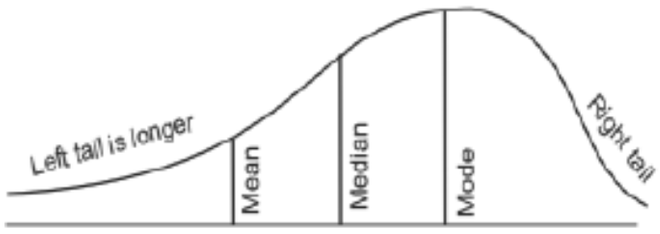
Skewness denotes lack of symmetry. The distribution is said to be skewed if mean, median and mode are not equal.

Skew Symmetrical Distribution

A distribution which is not symmetrical is said to be skew symmetrical distribution. In skew symmetrical distribution the left tail and the right tail are not of equal length. One tail will be longer than the other.

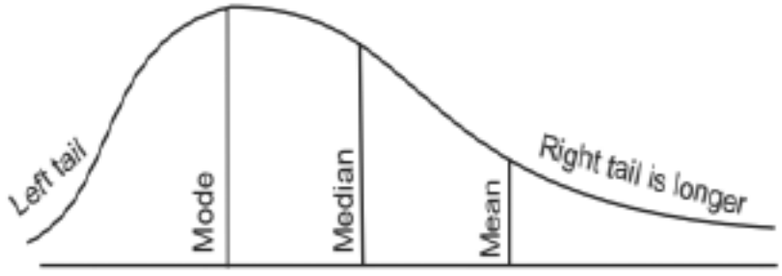
Negatively Skewed Distribution

In negatively skewed distribution, left tail of the curve is longer than the right tail.



Positively Skewed Distribution

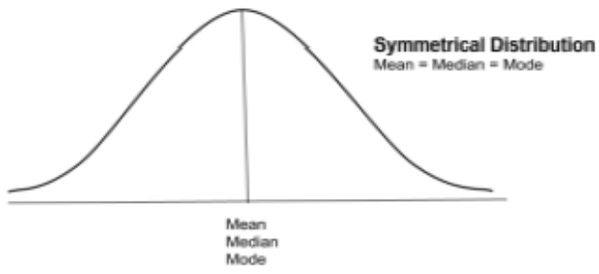
In positively skewed distribution, right tail of the curve is longer than the left tail.



Symmetric Distribution

In symmetric distribution, both the tails of the curve are same.

Notes



Definition of Skewness

A distribution is said to be **'skewed'** when the mean, median and mode fall at different points in the distribution and the balance is shifted to one side or the other – To left or right. It is denoted as S_k

Note:

(i) There is no skewness in the distribution if mean = mode = median

(ii) There is no skewness in the distribution if, third quartile – median = median – first quartile.

(iii) There is no skewness if the sum of the frequencies which are less than mode = sum of the frequencies which are greater than mode

(iv) There is no skewness if quartiles are equidistant from the median.

(v) The distribution is negatively skewed if mean is less than mode.

Types of Skewness

1. Fairly symmetrical
2. Positively skewed
3. Negatively skewed.

6.2.1 Measures of Skewness

Measure of skewness is known as the measure of symmetry. There are two types of measures of skewness.

1. **Absolute Measure:**
   - **i.** Skewness = (Mean – Mode)
   - **ii.** Skewness = (Mean – Median)
   - **iii.** Skewness = $(Q_3 - \text{Median}) - (\text{Median} - Q_1)$

Notes

2. Relative Measure: The relative measures of skewness are.

i. Karl Pearson's Coefficient of Skewness

ii. Bowley's Coefficient of Skewness.

6.2.2 Karl Pearson's Coefficient of Skewness

Karl Pearson's Coefficient of Skewness =

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

Or

$$= \frac{3\ (\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

It generally lies between –3 and 3.

If its value is zero then there is no skewness.

Note:

1. Distribution is symmetrical, that is, there is no skewness if $S_k = 0$

$$\frac{mean - mode}{standard\ deviation} = 0 \Rightarrow mean - mode = 0 \Rightarrow mean = mode$$

2. Distribution is negatively skewed if $S_k < 0$

$$\frac{mean - mode}{standard\ deviation} < 0 \Rightarrow mean - mode < 0 \Rightarrow mean < mode$$

3. Distribution is positively skewed if $S_k > 0$

$$\frac{mean - mode}{standard\ deviation} > 0 \Rightarrow mean - mode > 0 \Rightarrow mean > mode$$

Example 6: Compute the coefficient of Skewness from the following data:

X	6	7	8	9	10	11	12
F	3	6	9	13	8	5	4

Solution: Let A = 9

X	**f**	**d = x – 9**	**f d**	**f d^2**	**c.f.**
6	3	– 3	– 9	27	3
7	6	– 2	– 12	24	9
8	9	– 1	– 9	9	18
9	13	0	0	0	31
10	8	1	8	8	39

Notes

11	5	2	10	20	44
12	4	3	12	36	48
	$\Sigma f = 48$		$\Sigma fd = 0$	$\Sigma fd^2 = 124$	

$$\text{Mean} = A + \frac{\Sigma fd}{\Sigma f} = 9 + \frac{0}{48} = 9$$

Mode = value of x corresponding to maximum frequency (13) = 9

$$SD = \sqrt{\frac{\Sigma fd^2}{\Sigma f} - \left(\frac{\Sigma fd}{\Sigma f}\right)^2}$$

$$SD = \sqrt{\frac{124}{48} - \left(\frac{0}{48}\right)^2}$$

SD = 1.61

$$\text{Karl Pearson's Coefficient of Skewness} = \frac{mean - mode}{standard\ Deviation} = \frac{9-9}{1.61} = 0$$

Example 7: Calculate Karl Pearson's Coefficient of Skewness from the table given below:

Wages of Day	55 – 58	58 – 61	61 – 64	64 – 67	67 – 70
No. of Workers	12	17	23	18	11

Solution: Let A = 62.5

Wages of Day	**No. of Workers (*f*)**	***Mid Value* (*x*)**	**d = x – 62.5**	**fd**	**fd^2**	**c.f.**
55 – 58	12	56.5	– 6	– 72	432	12
58 – 61	17	59.5	– 3	– 51	153	29
61– 64	23	62.5	0	0	0	52
64 – 67	18	65.5	3	54	162	70
67 – 70	11	68.5	6	66	396	81
	$\Sigma f = 81$			$\Sigma fd = -3$	$\Sigma fd^2 = 1143$	

$$\text{Mean} = A + \frac{\Sigma fd}{\Sigma f} = 62.5 + \frac{-3}{81} = 62.46$$

Median class is 61 – 64

Notes

Hence, Median $= l + \dfrac{\left(\dfrac{N}{2} - cf\right)}{f} i = 61 + \dfrac{\left(\dfrac{81}{2} - 29\right)}{23}(3) = 62.5$

SD $= \sqrt{\dfrac{\Sigma fd^2}{\Sigma f} - \left(\dfrac{\Sigma fd}{\Sigma f}\right)^2} = \sqrt{\dfrac{1143}{81} - \left(\dfrac{-3}{81}\right)^2} = \sqrt{\dfrac{10286}{729}} = 3.76$

Karl Pearson's Coefficient of Skewness $= \dfrac{3(mean - median)}{SD} = -0.032$

6.2.3 Bowley's Coefficient of Skewness

Bowley's Coefficient of Skewness is based on the quartiles and median. A distribution is symmetrical if the distance between the first quartile and median is equal to the distance between the median and third quartile. It is defined as

Bowley's Coefficient of Skewness $= \dfrac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$

where, Q_3 is the third Quartile, Q_1 is the first Quartile.

Characteristics of Bowley's Coefficient of Skewness

1. If the distribution has open end or unequal class intervals then Pearson's Coefficient of Skewness cannot be calculated but Bowley's Coefficient of Skewness can be calculated.
2. Bowley's Coefficient of Skewness lies between – 1 and + 1.
3. Bowley's measure is calculated only from the continuous distribution with exclusive classes.

Limitations of Bowley's Coefficient of Skewness

1. It is based on the central 50% of the data and ignores the remaining 50% of the data on the extremes.
2. Bowley's formulae and Pearson's formulae cannot be compared. However, if the distribution is symmetrical then both coefficients are zero.

Example 8: From the following data find Bowley's Coefficient of Skewness: Difference of quartiles = 80, Mode = 60, Sum of the quartiles = 120 and Mean = 45

Notes

Solution: Here, we have

$$Q_3 + Q_1 = 120$$

$$Q_3 - Q_1 = 80$$

$$\text{Mode} = 60$$

$$\text{Mean} = 45$$

We know that

Mode = 3 Median – 2 Mean

60 = 3 Median – 2 (45)

Therefore, Median = 50

Now,

$$\text{Bowley's Coefficient of Skewness} = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$$

$$= \frac{120 - 2(50)}{80} = 0.25$$

Example 9: Calculate Bowley's Coefficient of Skewness from the data given below:

No. of Houses	0	1	2	3	4	5	6
No. of Air Conditioners	15	20	14	25	13	8	4

Solution:

No. of Houses (*x*)	**No. of Air Conditioners (*f*)**	**Cumulative Frequency**
0	15	15
1	20	35
2	14	49
3	25	74
4	13	87
5	8	95
6	4	99

Here N = 99

Q_1 = size of $\left(\frac{N+1}{4}\right)^{th}$ item = 1

Q_3 = size of $\left(\frac{3(N+1)}{4}\right)^{th}$ item = 4

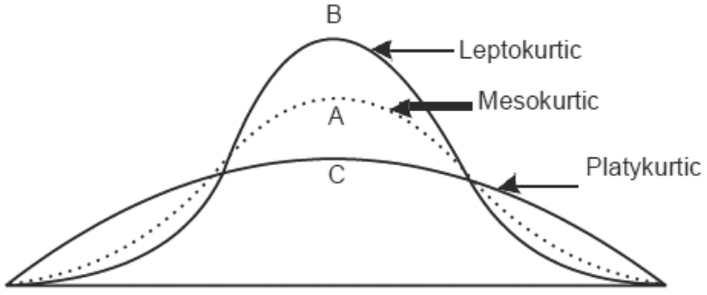
median = size of $\left(\frac{N+1}{2}\right)^{th}$ item = 3

Bowley's Coefficient of Skewness = $\frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$

$= \frac{4+1-2(3)}{4-1} = -0.33$

6.3 Kurtosis

It tells about the shape of a frequency distribution. It is a measure of the flatness or peakedness of the curve



The measure of Kurtosis is

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, \quad \gamma_2 = \beta_2 - 3$$

If β_2 = 3, i.e., $\gamma_2 = 0$ the curve is normal or mesokurtic.

β_2 > 3, i.e., $\gamma_2 > 0$ the curve is peaked than normal or leptokurtic.

β_2 < 3, i.e., $\gamma_2 < 0$ the curve is flat topped or platykurtic.

Note: Measure of Skewness in terms of moments is:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \quad \gamma_1 = +\sqrt{\beta_1}$$

Notes

Example 10: Find the relation between moment about the mean and moment about any arbitrary point. The first four moments of a distribution about the value 4 of the variate are –1.5, 17, –30 and 108. Calculate the first four moments about the mean and find β_1 and β_2.

Solution:

We have,

A = 4, $\mu_1' = -1.5$, $\mu_2' = 17$, $\mu_3' = -30$ and $\mu_1'^4 = 108$

Moments about the mean

$\mu_1 = 0$

$\mu_2 = \mu'_2 - (\mu'_1)^2 = 17 - (-1.5)^2 = 17 - 2.25 = 14.75$

$\mu_3 = \mu'_3 - 3\mu'_2\,\mu'_1 + 2\,(\mu'_1)^3$

$= -30 - 3\,(17)\,(-1.5) + 2\,(-1.5)^3 = -30 + 76.5 - 6.75$

$= 39.75$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$

$= 108 - 4\,(-30)\,(-1.5) + 6\,(17)\,(-1.5)^2 - 3\,(-1.5)^4$

$= 108 - 180 + 229.5 - 15.19 = 142.31$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{39.75^2}{14.75^3} = 0.4924$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{142.31}{14.75^2} = 0.6541$$

Example 11: 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: Let A = 4

x	**f**	**d = x – 4**	**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

Notes

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
	Σ f = 256		Σ fd = 0	Σ fd^2 = 512	Σ fd^3 = 0	Σ fd^4 = 2816

Moments about the point A = 4 are

$$\mu_1' = \frac{1}{N}\Sigma f_i(x_i - A)^1 = 0$$

$$\mu_2' = \frac{1}{N}\Sigma f_i(x_i - A)^2 = 2$$

$$\mu_3' = \frac{1}{N}\Sigma f_i(x_i - A)^3 = 0$$

$$\mu_4' = \frac{1}{N}\Sigma f_i(x_i - A)^4 = 11$$

Moments about mean are

$$\mu_2 = \mu_2' - \mu_1'^2 = 2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 11$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

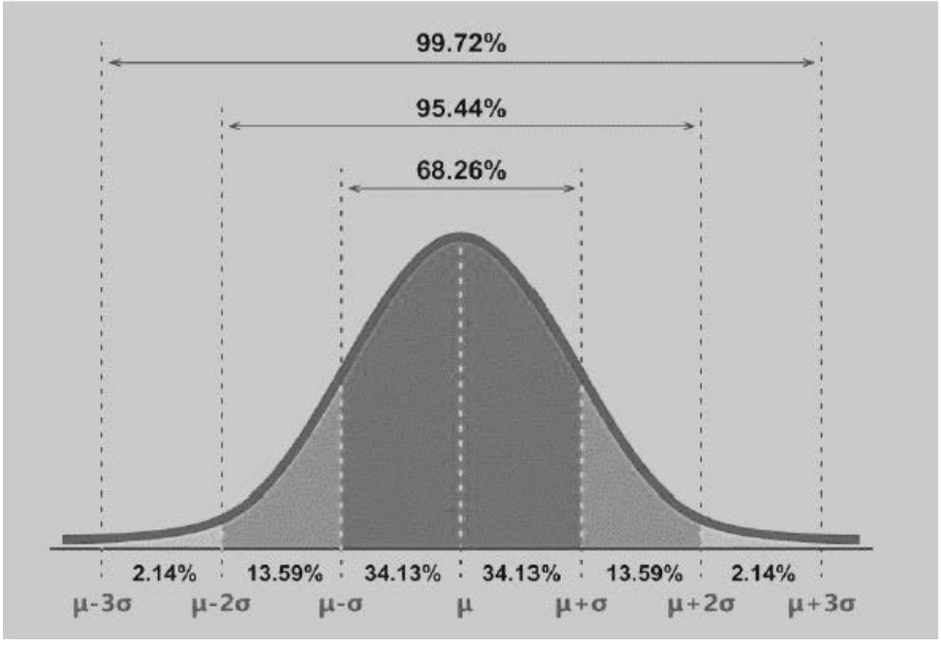
$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 2.75$$

6.4 Normal Curve

What is a Normal Distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve".

Notes



Graph Courtesy: Simply Psychology

Properties of the Normal Distribution

The normal distribution has several key features and properties that define it.

First, its mean (average), median (midpoint), and mode (most frequent observation) are all equal to one another. Moreover, these values all represent the peak, or highest point, of the distribution. The distribution then falls symmetrically around the mean, the width of which is defined by the standard deviation.

The Empirical Rule

For all normal distributions, 68.2% of the observations will appear within plus or minus one standard deviation of the mean; 95.4% of the observations will fall within +/- two standard deviations; and 99.7% within +/- three standard deviations. This fact is sometimes referred to as the "empirical rule," a heuristic that describes where most of the data in a normal distribution will appear.

Notes

The Formula for the Normal Distribution

The normal distribution follows the following formula. Note that only the values of the mean (μ) and standard deviation (σ) are necessary

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

where,

x = value of the variable or data being examined and f(x) the probability function

μ = the mean

σ = the standard deviation

6.5 Miscellaneous Questions

Example 12: Compute the first four moments about mean from the following data:

Mid Value	5	10	15	20	25	30	35
Frequencies	8	15	20	32	23	17	5

Solution: Let A = 20

x	**f**	**(x – 20)/5 = d**	**fd^1**	**$f(d)^2$**	**$f(d)^3$**	**$f(d)^4$**
5	8	-3	-24	72	-210	648
10	15	-2	-30	60	-124	240
15	20	-1	-20	20	-20	20
20	32	0	0	0	0	0
25	23	1	23	23	23	23
30	17	2	34	68	136	272
35	5	3	15	45	135	405
	$\Sigma f_i = 120$	0	$\Sigma f_i(d) = -2$	$\Sigma f_i(d)^2 = 288$	$\Sigma f_i(d)^3 = -62$	$\Sigma f_i(d)^4 = 1608$

Moments about arbitrary origin

$$\mu_1' = h \times \frac{1}{N}\sum_{i=1}^{7} f_i d^1 = 5 \times \frac{-2}{120} = -0.083$$

Notes

$$\mu_2' = h^2 \times \frac{1}{N}\sum_{i=1}^{7} f_i d^2 = 25 \times \frac{288}{120} = 60$$

$$\mu_3' = h^3 \times \frac{1}{N}\sum_{i=1}^{7} f_i d^3 = 125 \times \frac{-62}{120} = -64.58$$

$$\mu_4' = h^4 \times \frac{1}{N}\sum_{i=1}^{7} f_i d^4 = 625 \times \frac{1608}{120} = 8375$$

$\mu_1 = 0$

$\mu_2 = \mu_2' - \mu_1'^2 = 60 - (-0.083)^2 = 59.993$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$

$= -64.583 - 3(-0.083)(60) + 2(-0.083)^3$

$= -49.644$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 2\mu_1'^4$

$= 8375 - 4(-0.083)\ (-64.583) + 6(-0.082)^2(60) - 3(-0.083)^4$

$= 8251.08$

Example 13: Calculate the first four moments about the mean and comment on the nature of the distribution:

x	1	2	3	4	5	6	7	8	9
f	1	6	13	25	30	22	9	5	2

Solution:

x	**f**	**fx**	**(x –5) = d**	**fd**	**fd^2**	**fd^3**	**fd^4**
1	1	1	-4	-4	16	-64	256
2	6	12	-3	-18	54	-162	486
3	13	39	-2	-26	52	-104	208
4	25	100	-1	-25	25	-25	25
5	30	150	0	0	0	0	0
6	22	132	1	22	22	22	22
7	9	63	2	18	36	72	144
8	5	40	3	15	45	135	405
9	2	18	4	8	32	128	512
	$\Sigma f = 113$	$\Sigma fx = 555$		$\Sigma f_i(d) = -10$	$\Sigma f_i(d)^2 = 282$	$\Sigma f_i(d)^3 = 2$	$\Sigma f_i(d)^4 = 2058$

Notes

$$\text{Mean} = \frac{1}{N}\sum_{i=1}^{9} f_i x_i = \frac{555}{113} = 4.91$$

$$\mu_1' = \frac{1}{N}\sum_{i=1}^{9} f_i \left(x_i - 5\right)^1 = \frac{-10}{113}$$

$$\mu_2' = \frac{1}{N}\sum_{i=1}^{9} f_i \left(x_i - 5\right)^2 = \frac{282}{113}$$

$$\mu_3' = \frac{1}{N}\sum_{i=1}^{9} f_i \left(x_i - 5\right)^3 = \frac{2}{113}$$

$$\mu_4' = \frac{1}{N}\sum_{i=1}^{9} f_i \left(x_i - 5\right)^4 = \frac{2058}{113}$$

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = \frac{282}{113} - \left(\frac{-10}{113}\right)^2 = 2.496 - 0.0078 = 2.488$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \frac{2}{113} - 3\left(\frac{282}{113}\right)\left(\frac{-10}{113}\right) + 2\left(\frac{-10}{113}\right)^3$$

$$= 0.017699 + 0.662542 + 0.015663$$

$$= 0.69590$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 2\mu_1'^4$$

$$= \frac{2058}{113} - 4\left(\frac{2}{113}\right)\left(\frac{-10}{113}\right) + 6\left(\frac{282}{113}\right)\left(\frac{-10}{113}\right)^2 - 3\left(\frac{-10}{113}\right)^4$$

$$= 18.212 + 0.00626 + 0.111726 - 0.00018399$$

$$= 18.335$$

$$\beta_1 = skewness = \frac{\mu_3^2}{\mu_2^3} = \frac{0.69590^2}{2.488^3} = \frac{0.484277}{15.40108} = 0.03144$$

$$\beta_2 = kurtosis = \frac{\mu_4}{\mu_2^2} = \frac{18.335}{2.488^2} = \frac{18.335}{6.19} = 2.96$$

Notes

Since $\beta_1 > 0$ the distribution is positively skewed

Since $\beta_2 < 3$ the distribution is platykurtic.

Example 14: For a group of 20 items, $\Sigma X = 1452$, $\Sigma X^2 = 144280$ and mode = 63.7. Find the Pearsonian coefficient of skewness.

Solution:

Pearsonian coefficient of skewness = $S_k = \frac{Mean - Mode}{S.D.}$

Mean = $\frac{\Sigma X}{N} = \frac{1452}{20} = 72.6$

S.D. = $\sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2} = \sqrt{\frac{144280}{20} - (72.6)^2} = 44.08$

Hence,

$$S_k = \frac{72.6 - 63.7}{44.08} = 0.202$$

Example 15: Calculate the Karl Pearson's coefficient of skewness from the following data:

Size (x)	3.5	4.5	5.5	6.5	7.5	8.5	9.5
Frequency (f)	3	7	22	60	85	32	8

Solution:

x	**f**	**x – 6.5 = d**	**fd**	**fd^2**
3.5	3	-3	-9	27
4.5	7	-2	-14	28
5.5	22	-1	-22	22
6.5	60	0	0	0
7.5	85	1	85	85
8.5	32	2	64	128
9.5	8	3	24	72
	Σ f = 217	0	Σ fd = 128	Σ f d^2 = 362

Notes

Karl Pearson's coefficient of skewness = $S_k = \frac{Mean - Mode}{S.D.}$

Here, A = 6.5, N = 217

$$\text{Mean} = A + \frac{\Sigma fd}{N} = 6.5 + \frac{128}{217} = 6.5 + 0.59 = 7.09$$

$$\text{S.D.} = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{362}{217} - \left(\frac{128}{217}\right)^2} = 1.149$$

By inspection, mode is 7.5, the size with maximum frequency

Hence,

$$S_k = \frac{7.09 - 7.5}{1.149} = -0.357$$

Since $S_k < 0$ hence distribution is negatively skewed.

Example 16: Compute Bowley's Coefficient of Skewness from the data given below:

Marks	35-36	36-37	37-38	38-39	40-41	41-42	42-43
No. of Students	14	20	42	54	45	21	8

Solution:

Marks	**No. of Persons (f)**	**Cumulative Frequency (c.f.)**
35 – 36	14	14
36 – 37	20	34
37 – 38	42	76
38 – 39	54	130
40 – 41	45	175
41 – 42	21	196
42 – 43	8	204

Here N = 204

Q_1 = size of $\left(\frac{N}{4}\right)^{th}$ item = 51^{th} item. Hence Q_1 lies in the class 37 – 38

Notes

$$Q_1 = l + \frac{\left(\frac{N}{4} - cf\right)}{f} \times i = 37 + \frac{\left(\frac{204}{4} - 34\right)}{42}(1)$$

$$= 37.405$$

Q_3 = size of $\left(\frac{3N}{4}\right)^{th}$ item = 153th item.

Hence Q_3 will lie in the class 40 – 41

$$Q_3 = l + \frac{\left(\frac{3N}{4} - cf\right)}{f} \times i = 40 + \frac{\left(\frac{3(204)}{4} - 130\right)}{45}(1)$$

$$= 40 + 0.511$$

$$= 40.511$$

Median = size of $\left(\frac{N}{2}\right)^{th}$ item = 102th item. Hence median will lie in the class 38 – 39.

$$Median = l + \frac{\left(\frac{N}{2} - cf\right)}{f} \times i = 38 + \frac{\left(\frac{204}{2} - 76\right)}{54}(1)$$

$$= 38 + 0.48$$

$$= 38.48$$

Now,

Bowley's Coefficient of Skewness = $\frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$

$$= \frac{40.511 + 37.405 - 2(38.48)}{40.511 - 37.405} = 0.307$$

Example 17: Compute Bowley's Coefficient of Skewness from the data given below:

Income (Rs.)	Below 200	200-400	400-600	600-800	800-1000	Above 1000
No. of Persons	25	40	80	75	20	16

Notes

Solution:

Income (Rs.)	No. of Air Persons (f)	Cumulative Frequency (c.f.)
Below 200	25	25
200-400	40	65
400-600	80	145
600-800	75	220
800-1000	20	240
Above 1000	16	256

Here N = 256

Q_1 = size of $\left(\frac{N}{4}\right)^{th}$ item = 64th item. Hence Q_1 lies in the class 200-400

$$Q_1 = l + \frac{\left(\frac{N}{4} - cf\right)}{f} \times i = 200 + \frac{\left(\frac{256}{4} - 25\right)}{40}(200)$$

$$= 200 + 195 = 395$$

Q_3 = size of $\left(\frac{3N}{4}\right)^{th}$ item = 192th item. Hence Q_3 will lie in the class 600-800

$$Q_3 = l + \frac{\left(\frac{3N}{4} - cf\right)}{f} \times i = 600 + \frac{\left(\frac{3(256)}{4} - 145\right)}{75}(200)$$

$$= 600 + 125.33 = 725.33$$

median = size of $\left(\frac{N}{2}\right)^{th}$ item = 128th item. Hence median will lie in the class 400 – 600.

$$Median = l + \frac{\left(\frac{N}{2} - cf\right)}{f} \times i = 400 + \frac{\left(\frac{256}{2} - 65\right)}{80}(200)$$

$$= 400 + 157.5 = 557.5$$

Notes

Now,

Bowley's Coefficient of Skewness = $\frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$

$= \frac{725.33+395-2(557.5)}{725.33-395} = 0.016$

Example 18: For a distribution, mean = 10, variance = 16, $\gamma_1 = +1$, $\beta_2 = 4$. Obtain the first four moments about the origin.

Solution:

We have,

Mean = 10, variance = 16, $\gamma_1 = +1$, $\beta_2 = 4$

The first moment about zero = μ'_1 = mean = 10

The second moment about mean = μ_2 = variance = 16

Second moment about zero = $\mu'_2 = \mu_2 + (\mu'_1)^2 = 16 + (10)^2 = 116$

Now

$$\gamma_1 = +\sqrt{\beta_1}$$

Or $\gamma_1^2 = \beta_1$

$1^2 = \beta_1$

Or $\beta_1 = 1$

Also, $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

$1 = \frac{\mu_3^2}{16^3}$

Or $\mu_3^2 = 4096$

$\Rightarrow \mu_3 = 64$

Hence, third moment about origin is $\mu'_3 = \mu_3 + 3\mu'_2\,\mu'_1 - 2(\mu'_1)^3$

$= 64 + 3(116)(10) - 2(10)^3 = 64 + 3480 - 2000 = 1544$

We have $\beta_2 = 4$

Therefore, $\beta_2 = \frac{\mu_4}{\mu_2^2} = 4$

Notes

Or $\mu_4 = \mu_2^2 = 4(16)^2 = 1024$

Now, fourth moment about zero = μ_4'

$$= \mu_4 + 4\mu_3'\mu_1' - 6\mu_2'\mu_1'^2 + 3\mu_1'^4$$

$$= 1024 + 4(1544)(10) - 6(116)(10)^2 + 3(10)^4$$

$$= 1024 + 61760 - 69600 + 30000$$

$$= 23184$$

Example 19: The daily expenditure of 100 families is given below:

Daily Expenditure	0 - 20	20 - 40	40 - 60	60 - 80	80 - 100
No. of Families	13	?	27	?	16

If the mode of the distribution is 44, calculate the Karl Pearson coefficient of skewness

Solution: First, we calculate the missing frequencies

Daily Expenditure	**Mid Value (x)**	**No. of Families (f)**
0 – 20	10	13
20 – 40	30	x
40 – 60	50	27
60 – 80	70	y
80 – 100	90	16
		N = 100

Since mode is 44

Therefore, modal class is 40 – 60

Thus, $f_1 = 27$, $f_0 = x$, $f_2 = y$, $l = 40$, $h = 20$

Now we know mode is given by

$$M_o = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

$$44 = 40 + \frac{27 - x}{2(27) - x - y} \times 20$$

$$\frac{44 - 40}{20} = \frac{27 - x}{54 - x - y}$$

Notes

$$\frac{1}{5}=\frac{27-x}{54-x-y}$$

Or $54 - x - y = 5(27 - x)$

Or $54 - x - y = 135 - 5x$

Or $4x - y = 81$ ------ *

Also total frequency = N = 100

Thus, $13 + x + 27 + y + 16 = 100$

Or $x + y = 44$ ------ **

Solving * and **, we get

$x = 25$ and $y = 19$

Now, we calculate Karl Pearson coefficient of skewness

Daily Expenditure	Mid Value (x)	f	(x – 50)/20 = d	fd	fd^2
0 - 20	10	13	-2	-26	52
20 – 40	30	25	-1	-25	25
40 – 60	50	27	0	0	0
60 – 80	70	19	1	19	19
80 – 100	90	16	2	32	64
		$\Sigma f = 100$		$\Sigma fd = 0$	$\Sigma fd^2 = 160$

Karl Pearson's coefficient of skewness = $S_k = \frac{Mean - Mode}{S.D.}$

Here, A = 6.5, N = 100

$$\text{Mean} = A + \frac{\Sigma fd}{N} = 50 + \frac{0}{100} \times 20 = 50 + 0 = 50$$

$$\text{S.D.} = i \times \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = 20 \times \sqrt{\frac{160}{100} - \left(\frac{0}{100}\right)^2} = 25.2$$

Mode = 44

Hence,

$$S_k = \frac{50 - 44}{25.2} = 0.238$$

Since $S_k > 0$ hence distribution is positively skewed.

Notes

6.6 Exercise

1. Calculate first four moments about the mean, for the following individual series:

5	5	5	5	5	5

2. Find the first four moments about the mean of the following series:

1	3	7	9	10

3. If the first four moments of a distribution about the value 5 are equal to –4, 22, –117 and 560. Determine the corresponding moments:

About the (i) mean, and (ii) about zero

4. Compute first four moments of the data 3, 5, 7, 9 about the mean. Also, compute the first four moments about the point

5. Calculate Karl Pearson's Coefficient of Skewness from the data given below:

1. S.D. = 6.5, mean = 29.6, mode = 27.52.

2. Mean = 100, Variance = 35, Median = 99.61.

3. Mean = 45, Median = 48, S.D. = 22.5.

6. Find the Karl Pearson's Coefficient of Skewness for the following:

Years Under	10	20	30	40	50	60
No. of Persons	15	32	51	78	97	109

7. Calculate Karl Pearson's Coefficient of Skewness from the following data:

Cost per Item (in Rs.)	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5
No. of Items	35	40	48	100	125	87	43	22

8. The data for a distribution is given below Q_1 = 8.6, Median = 12.3, Q_3 = 14.04. Calculate Bowley's Coefficient of Skewness.

9. In a routine checkup the weights of the students of government school were noted as follows:

Weights	55–58	58–61	61–64	64–67	67–70
No. of Students	12	17	23	18	11

Notes

Calculate the Bowley's Coefficient of Skewness.

10. Calculate the first four moments about the mean and also the value of skewness and Kurtosis from the following table.

x	25	35	45	55	65	75	85
f	5	14	20	25	17	11	8

11. Calculate the first four moments about the mean from the following data:

x	0	1	2	3	4	5	6	7	8
f	5	10	15	20	25	20	15	10	5

Also, calculate the values of skewness and Kurtosis and comment on the nature of the distribution.

12. Analyse the frequency distribution by the method of moments:

x	2	3	4	5	6
f	1	3	7	3	1

13. In a distribution the difference of two quartiles is 2.63, their sum is 62.55 and the median is 36.03. Find the coefficient of skewness.

14. From the information given below calculate Karl Pearson's and Bowley's coefficient of skewness

Mean =150; Median = 142; S.D. = 30; Third Quartile = 195; First Quartile = 62

15. Find the coefficient of skewness from the following information:

Difference of two quartiles = 8, Mode = 11, Mean = 8, sum of two quartiles = 22

16. Pearson's coefficient of skewness for a data distribution is 0.5 and coefficient of variation is 40%. Its mode is 80. Find the mean and the median of the distribution.

17. The following data are given to an economist for the purpose of economic analysis. The data refer to the length of a sample of Good Year Tyres. Do you think the distribution is symmetric and platykurtic?

$$N = 100, \Sigma f d_x = 50, \quad \Sigma f d_x^2 = 1967.2,$$
$$\Sigma f d_x^3 = 2925.8, \quad \Sigma f d_x^4 = 86650.2$$

Notes

18. Given the following information, find the first four central moments

$$N = 10, \Sigma f d_x = -100, \quad \Sigma f d_x^2 = 400,$$
$$\Sigma f d_x^3 = -1000, \quad \Sigma f d_x^4 = 5000$$

19. Calculate Bowley's coefficient of skewness from the following data and comment on the value:

Age (Yrs.)	No. of Employees	Age (Yrs.)	No. of Employees
Below 20	13	35 – 40	72
20 – 25	29	40 – 45	94
25 – 30	46	45 – 50	45
30 – 35	60	50 – 55	21

20. The mean, mode and Quartile deviation of a distribution are 42, 36 and 15 respectively. If its Bowley's coefficient of skewness is 1/3, find the values of the two quartiles.

21. Find the four moments about mean from the following data. Also decide whether it is a platykurtic distribution.

Size of the Item	1	2	3	4	5
Frequency	2	3	5	4	1

22. For a distribution, the mean is 10, standard deviation is 4, $\beta_1 = 1$, and $\beta_2 = 4$. Obtain the first four moments about 4. Comment upon the nature of the distribution.

Department of Distance and Continuing Education
Campus of Open Learning, School of Open Learning, University of Delhi