INTRODUCTION TO STATISTICS

Extended with exercises

Statistics is the knowledge of numbers and how to play with them to make our point stand out, throughout this book you will learn how to use numbers in decision making as a manager and in policy making as a technocrat.

Eng. A. S. Noorzad
Qais_Mohammadi

Basic Statistics

I am proud of you Dad; however, I would not dare to tell it to your face.
This book is dedicated to you!

About the authors

Eng. Abdul Sami Noorzad is a lecturer at Karwan University. He specializes in teaching Mathematics and Physics. He received his M.Sc. Mathematics from JMI University New Delhi and studies Quantum Field Theory in the Department of Physics and Astrophysics, Delhi University. Currently he is busy on his research at Universitat Duisburg-Essen. He has authored more 60 books on Mathematics, Physics and Statistics. He has also authored and presented dozens of academic articles on mathematics and physics around the world. Some of his books are "Complex Numbers", "A Review of Trigonometry", "Some Facts about Mathematics", Methods of Solving Derivatives", Table of Integrals" and many more. Mr. Noorzad lives in Kabul with his family.

Qais Mohammadi is Lecturer of Economics at Karwan University, Kabul Afghanistan where he specializes in teaching introductory Statistics and Economics. He received his BS in Economics and Statistics from Kabul University (2011) and his Masters in Economics from Punjab University, Chandigarh India (2014). Meanwhile he continues his Masters in Public Administration and Policy in Kabul University (2016).
His research is primarily in the areas of development economics. He has authored dozens of scholarly articles on the economy of Afghanistan such as his recent research on "Chabahar The Decay of Afghanistan's Geo-Strategic Location", "Exchange rate of Afghani", "Free land for a better Afghanistan, decreased wages and increased exports" and many other articles published in Karwan University Research Magazine. On the other hand, this book is a milestone for future publications.
Qais Mohammadi teaches courses in each of these fields (Microeconomics, Macroeconomics, Development Economics, International Economics, Statistics, Econometrics and Research Methodology). He lives in Ahmad Shah Baba Mena Kabul, Afghanistan.

Brief Contents

Contents

Contents

Preface

As a lecturer of statistics and economics I have seen the least use of numbers in the afghan society, bureaucrats and scholars, which has always wondered me why? After asking and talking to a lot of students and colleagues I found one of the reasons being inability of interpreting data and the other being, lower attention given to the subject both in schools and universities. However, the use of statistics in day to day activities is practiced by almost all the people, they simply don't know it.

Keeping all this in view, the present book has been written with two clear objectives, viz., (i) to encourage the knowledge of using statistics among students and academicians, (ii) to help students and other interested parties in interpretation of data.

In order to achieve the above mentioned objective this book is compiled in seven chapters. The first chapter focuses attention on introduction to data, second and third chapters are focused on illustration of data, while the remaining chapters discuss central tendency measurements and more technical aspects of statistics.

Acknowledgement

Our deep appreciation and heartfelt thanks to the following personalities who has helped us shape this text.

Sayed Javid Andish founder of Karwan University who has tolerated burdensome problems in availing higher education for a developed Afghanistan.

Tawos Mohammadi Lecturer at Kabul Education University, Language and Literature Faculty, English Dept. - Linguistic support.

Students of BBA Department at Karwan University

1

Introduction to Statistics: Data handling

Introduction

You must have observed your teacher recording the attendance of students in your class every day, or recording marks obtained by you after every test or examination. Similarly, you must have also seen a cricket score board. Two score boards have been illustrated here:

Name of the ballers	Overs	Overs Made	Runs made	Wickets taken
A	10	2	40	3
B	10	1	30	2
C	10	2	20	1
D	10	1	50	4

Name of the batsmen	Runs	Balls faced	Time (in min)
E	45	62	75
F	55	70	81
G	37	53	67
H	22	41	55

You know that in a game of cricket the information recorded is not simply about who won and who lost. In the score board, you will also find some equally important information about the game. For instance, you may find out the time taken and number of balls against the highest runs scored. Similarly, in your day to day life, you must have seen several kinds of tables consisting of numbers, figures, names etc. These tables provide “Data”. Data is the collection of numbers gathered to give some information.

Gathering data

Collecting data involves two key decisions. The first refers to what to measure. Unfortunately, it is not necessarily the case that the easiest to measure variable is the most relevant for the specific problem in hand. The second relates to how to obtain the data. Sometimes gathering data is costless, e.g., a simple matter of internet downloading. However, there are many situations in which one must take a more active approach and constitute data set from the scratch.

Data gathering normally involves either sampling or experimentation. Albeit the latter is less common in social sciences, one should always have in mind that there is no need for a lab to run an experiment. There is pretty of room for experimentation within organizations. And we are not speaking exclusively about research and development. For instance, we could envision a sales competition to test how salespeople react to different levels of performance incentive. This is just one example of a key driver to improve quality of products and processes.

Sampling is a much more natural approach in social sciences. It is easy to appreciate that it is sometimes too costly, if not impossible, to gather universal data and hence it makes sense to restrict attention to a representative sample of the population. For instance, while census data are available only every 5 to 10 years due to the enormous cost/effort that it involves, there are several household and business surveys at the annual, quarterly, monthly, and sometimes even weekly frequency.

Recording data

Let us take the example of an elementary school class (say class 6), which is preparing to go for a picnic. The teacher asked the students to give their choice of fruits out of banana, apple, orange or melon. Yama is asked to prepare the list. He prepared a list of all the children and wrote the choice of fruit against each name. This list would help the teacher to distribute fruits according to the choice.

Student	Fruit	Student	Fruit
Ghulam	Apple	Farida	Banana
Jamshid	Melon	Zuhra	Apple
Raza	Orange	Zainab	Banana
Aslam	Apple	Farida	Orange
Salman	Banana	Friba	Melon
Hameed	Orange	Nazia	Apple
Najib	Melon	Zarlasht	Banana

If the teacher wants to know the number of bananas required for the class, He has to read the names in the list one by one and count the total number of bananas required. To know the number of apples, melons and oranges separately he has to repeat the same process for each of these fruits. How tedious and time consuming it is! It might become more tedious if the list has, say 50 students.

So Yama writes only the name of these fruits one by one like, banana, apple, melon, orange, apple, banana, orange, melon ,banana, banana, apple, banana, apple, banana, orange, melon, apple, banana, melon, banana.

Do you think this makes the teacher's work easier? He still has to count the fruits in the list one by one as he did earlier.

Zarlasht has another idea. She makes four squares on the floor. Every square is kept for fruit of one kind only. She asks the students to put one pebble in the square which matches their choices, i.e. a student opting for banana will put a pebble in the square marked for banana and so on.

By counting the pebbles in each square, Zarlasht can quickly tell the number of each kind of fruit required. She can get the required information quickly by systematically placing the pebbles in different squares.

Try to perform this activity for 40 students and with names of any four fruits. Instead of pebbles you can also use bottle caps or some other token.

Organizing data

To get the same information which Zarlasht got, Ahmad needs only a pen and a paper. He does not need pebbles. He also does not ask students to come and place the pebbles. He prepares the following table.

Banana	√	√	√	√	√	√	√	√	8
Orange	√	√	√						3
Apple	√	√	√	√	√				5
Melon	√	√	√	√					4

Do you understand Ahmad's table?

What does one $(\surd)$ mark indicate?

Four students preferred melon. How many $(\surd)$ marks are there against melon?

How many students where there in the class? Find all this information.

Discuss about these methods. Which is the best? Why? Which method is more useful, when information from a much larger data is required?

Example 1: A teacher wants to know the choice of food of each student as part of the mid-day meal program. The teacher assigns the task of collecting this information to Firoza. Firoza does so using a paper and a pencil. After arranging the choices in a column, she puts against a choice of food one slush (/) mark for every student making that choice.

Choice	Number of students
Qhaboli	/// /// /// //
Kabab	/// /// /// ///
Bolani	/// //

Raza, after seeing the table suggested a better method to count the students. He asked Firoza to organize the marks (I) in groups of ten as shown below:

Choice	Tally marks	Number of students
Qhaboli	(IIIIIIIIII) IIIIIII	17
Kabab	(IIIIIIIIII) III	13
Bolani	(IIIIIIIIII) (IIIIIIIIII)	20

Aslam made it simpler by asking him to make groups of five instead of ten, as shown below:

Choice	Tally marks	Number of students
Qhaboli	(IIIII) (IIIII) (IIIII) II	17
Kabab	(IIIII) (IIIII) III	13
Qhaboli and Kabab	(IIIII) (IIIII) (IIIII) (IIIII)	20

Teacher suggested that the fifth mark in a group of five marks should be used as a cross.

Example 2: Nazia is asked to collect data for the size of shoes of students in her class. Her findings are recorded in the manner shown below:

5	4	7	5	6	7	6	5	6	6	5	4	5	6	8	7	4	6	5	6	4	6	5
7	6	7	5	7	6	4	8	7														

Jamshid wanted to know

i. The size of shoes worn by the maximum number of students.

ii. The size of shoes worn by the minimum number of students.

Can you find this information?

Nazia prepared a table using tally marks.

Shoe size	Tally marks	Number of students
4	𝍸	5
5	𝍸 III	8
6	𝍸 𝍸	10
7	𝍸 II	7
8	II	2

Now the questions asked earlier could be answered easily. You may also do such activities in your class using tally marks.

DO THIS

1. Collect information regarding the number of family members of your classmates and represent it in the form of a table. Find to which category most students belong.

Number of family members	Tally marks	Number of students with that many family members

Make a table and enter the data using tally marks. Find the number that appeared

a. The minimum number of times?
b. The maximum number of times?
c. Same number of times?

Data description

The first step of data analysis is to summarize the data by drawing plots and charts as well as by computing some descriptive statistics. These tools essentially aim to provide a better understanding of how frequent the distinct data values are, and of how much variability there is around a typical value in the data.

Data Distribution

It is well-known that a picture tells more than a million words. The same applies to any serious data analysis for graphs are certainly among the best and most convenient data

descriptors. We start with a very simple, though extremely useful, type of data plot that reveals the frequency at which any given data value appears in the sample. A frequency table reports the number of times that a given observations occurs or, if based on relative terms, the frequency of that value divided by the number of observations in the sample.

Pictograph

A cupboard has five compartments. In each compartment a row of books is arranged.
The details are indicated in the adjoining table:

Rows			**Number**	**Of**	**Books**	**@**	**= 1**	**book**
Row 1	@	@	@	@				
Row 2	@	@	@	@	@			
Row 3	@	@						
Row 4	@	@	@	@	@	@	@	@
Row 5	@	@	@					

Which row has the greatest number of books? Which row has the least number of books? Is there any row which does not have books?
You can answer these questions by just studying the diagram. The picture visually helps you to understand the data. It is a pictograph.
A pictograph represents data through pictures of objects. It helps answer the questions on the data at a glance.

DO THIS

Pictographs are often used by dailies and magazines to attract reader's attention.
Collect one or two such published pictographs and display them in your class. Try to understand what they say.
It requires some practice to understand the information given by a pictograph.

Interpretation of a pictograph

Example 3: The following pictograph shows the number of absentees in a class of 30 students during the previous week:

Days			**Number of absentees**			☺ = 1 **Absentee**		
Mon	☺	☺	☺	☺	☺			
Tue	☺	☺	☺	☺				
Wed	☺	☺						
Thu								
Fri	☺							
Sat	☺	☺	☺	☺	☺	☺	☺	☺

a. On which day were the maximum number of students absent?
b. Which day had full attendance?
c. What was the total number of absentees in that week?

SOLUTION

a. Maximum absentees were on Saturday. (There are 8 pictures in the row for Saturday; on all other days, the number of pictures are less).
b. Against Thursday, there is no picture, i. e. no one is absent. Thus, on that day the class had full attendance.
c. There are 20 pictures in all. So, the total number of absentees in that week was20.

Example 4: the colors of fridges preferred by people living in a locality are shown by the following pictograph:

Colors		**Number of people**			☺ = 10 people, ● = 5 people	
Blue	☺	☺	☺	☺	☺	
Green	☺	☺	☺			
Red	☺	☺	☺	☺	☺	●
White	☺	☺				

a. Find the number of people preferring blue color.
b. How many people liked red color?

Solution:

a. Blue color is preferred by 40 people. (☺=10, so 4 pictures indicate$4 \times 10\ people$).

b. Deciding the number of people liking red color needs more care. For 5 complete pictures, we get $5 \times 10 = 50$ people. For the last black picture, we may roughly take it as 5. So, number of people preferring red color is nearly 55.

THINK, DISCUSS AND WRITE

In the above example, the number of people who like red color was taken as $50 + 5$. If your friend wishes to take it as$50 + 8$, is it acceptable?

Example 5: A survey was carried out on 30 student of a class in an elementary school. Data about the different modes of transport used by them to travel to school was displayed as pictograph.

What can you conclude from the pictograph?

Modes of travelling				**Number of students**				**☺ - One Student**			
Private car	☺	☺	☺	☺							
Public bus	☺	☺	☺	☺	☺						
School Bus	☺	☺	☺	☺	☺	☺	☺	☺	☺	☺	☺
Cycle	☺	☺	☺								
Walking	☺	☺	☺	☺	☺	☺	☺				

Solution: From the pictograph we find that:

a. The number of students coming by private cars is 4.
b. Maximum number of students use the school bus. This is the most popular way.
c. Cycle is used by only three students.
d. The number of students using the other modes can be similarly found.

Example 6: Following is the pictograph of the number of wrist watches manufactured by a factory in a particular week.

Days	**Number of wrist watches manufactured**					**Ω = 100 watch, ◌ = 50**		
Mon	Ω	Ω	Ω	Ω	Ω	Ω		
Tue	Ω	Ω	Ω	Ω	Ω	Ω	Ω	◌
Wed	Ω	Ω	Ω	Ω	Ω	Ω	◌	
Thu	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω
Fri	Ω	Ω	Ω	Ω	Ω	Ω		
Sat	Ω	Ω	Ω	Ω	Ω	◌		

a. On which day were the least number of wrist watches manufactured?
b. On which day were the maximum number of wrist watches manufactured?

c. Find out the approximate number of wrist watches manufactured in the particular week?

We can complete the following table and find the answers.

Days	Number of wrist watches manufactured
Mon	600
Tue	More than 700 and less than 800
Wed	……
Thu	……..
Fri	……..
Sat	……..

EXERCISE

1. In a math test, the following marks were obtained by 40 students. Arrange these marks in a table using tally marks.

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

a. Find how many students obtained marks equal to or more than 7.
b. How many students obtained marks below 4?

2. Following is the choice of wrist watches of 30 students of class BBA.

Omega, Rado, Omega, Casio, Omega, Omega, Omega, Swatch, Swatch, Omega, Rado, Rado, Seko, Casio, Omega, Rado, Omega, Casio, Omega, Omega, Omega, Swatch, Swatch, Omega, Rado, Omega, Casio, Omega, Omega, Casio.

a. Arrange the names of watches in a table using tally marks.
b. Which watch is preferred most of the student?

3. Abdul Samad threw a dice 40 times, below the dots noted are shown:

1	3	5	6	6	3	5	4	1	6	2	5	3	4	6	1	5	5	6	1	1	2	2		
3	5	2	4	5	5	6	5	1	6	2	3	5	2	4	1	5								

Make a table and enter the data using tally marks.
Find the number that appeared:

a. The minimum number of times.
b. The maximum number of times.
c. Find those numbers that appear an equal number of times.

4. Following pictograph shows the number of tractors in five villages.

Villages	Number of tractors	= One Tractor
Village A	5	1
Village B	4	0
Village C	5	3
Village D	3	0
Village E	4	1

Observe the pictograph and answer the following questions.

a. Which village has the minimum number of tractors?
b. Which village has the maximum number of tractors?
c. How many more tractors village C has as compared to village B.
d. What is the total number of tractors in all the five villages?

5. The number of girl students in each class of Karwan university is depicted by the pictograph:

Classes	Number of girl students	☺ = 4 girls, ☻ = 2 girls
I	3	3
II	3	1
III	3	2
IV	3	1
V	2	0
VI	3	1
VII	3	0
VIII	1	0

Observe this pictograph and answer the following questions:

a. Which class has the minimum number of girl students?
b. Is the number of girls in class VI less than the number of girls in class V?
c. How many girls are there in class VII?

6. The sale of electronic bulbs on different days of a week is shown below:

Days	Number of electric bulbs	☼ = 2 Bulb
Mon	6	0
Tue	6	2
Wed	4	0
Thu	5	0
Fri	6	1
Sat	4	0

Sun	☼	☼	☼	☼	☼	☼	☼	☼	☼

What can we conclude from the said pictograph?

7. In a village six fruit merchants sold the following number of fruit baskets in a particular season.

Name of fruit merchants	**Number of fruit baskets**							**₪= 100 box, ┌ = 50**		
Rahim	₪	₪	₪	₪						
Ahmad	₪	₪	₪	₪	₪	┌				
Poya	₪	₪	₪	₪	₪	₪	₪			
Akmal	₪	₪	₪	₪	₪	₪	₪	₪	₪	┌
Akbar	₪	₪	₪	₪	₪	₪	₪	₪		
Wahid	₪	₪	₪	₪	┌					

Observe this pictograph and answer the following questions:

a. Which merchant sold the maximum number of baskets?
b. How many fruit baskets were sold by Poya?
c. The merchants who have sold 600 or more number of baskets are planning to buy a go down for the next season. Can you name them?

Drawing a pictograph

Drawing a pictograph is interesting. But sometimes, a symbol like ₪ (which was used in one of the previous examples) may represent multiple units and may be difficult to draw. Instead of it we can use simpler symbols. If ☺ represents say 5 students, how will you represent, say 4 or 3 students?

We can solve such a situation by making an assumption that ☺ represent 5 students, ☻ represents 4 students, O represents 3 students, o represents 2 students, and ◦ represents 1 student, and then start the task of representation.

Example 7: The following are the details of number of students present in a class of 30 during a week. Represent it by a pictograph.

Days	**Number of students present**
Monday	24
Tuesday	26
Wednesday	28
Thursday	30
Friday	29
Saturday	22

Solution: with the assumptions we have made earlier, 24 may be represented by ☺☺☺☺☻, 26 may represented by ☺☺☺☺☺◦, and so on. Thus, the pictograph would be

Days	Number of students present
Monday	☺☺☺☺☻
Tuesday	☺☺☺☺☺◦
Wednesday	☺☺☺☺☺o
Thursday	☺☺☺☺☺☺
Friday	☺☺☺☺☺☻
Saturday	☺☺☺☺o

We had some sort of agreement over how to represent "less than 5" by a picture. Such a sort of splitting the pictures may not be always possible. In such cases what shall we do? Study the following example.

Example 8: The following are the number of electric bulbs purchased for a lodging house during the first four months of a year.

Months	Number of bulbs
January	20
February	26
March	30
April	34

Represent the details by a pictograph.

Solution: Picturing for January and March is not difficult. But representing 26 and 34 with the pictures is not easy.

We may round off 26 to nearest 5 i. e. to 25 and 34 to 35. We then show two and a half bulbs for February and three and a half for April.

Months		Let ☼ represent 10 bulb, ○ = 5 bulbs		
January	☼	☼		
February	☼	☼	○	
March	☼	☼	☼	
April	☼	☼	☼	○

Chapter Exercise

1. Total number of animals in five villages are as follows:

Village A	80
Village B	120
Village C	90
Village D	40
Village E	60

Prepare a pictograph of these animals using one symbol ◙ to represent 10 animals and answer the following questions:

a. How many symbols represent animals of village E?
b. Which village has the maximum number of animals?
c. Which village has more animals, village A or village C?

2. Total number of students of a school in different years is shown in the following table

Years	Number of students
1996	400
1998	535
2000	472
2002	600
2004	623

A. Prepare a pictograph of students using one symbol ☺ to represent 100 students and answer the following questions:

a. How many symbols represent total number of students in the year 2002?
b. How many symbols represent total number of students for the year 1998?

B. Prepare another pictograph of students using any other symbol each representing 50 students. Which pictograph do you find more informative?

2

Two variable Graphs

Frequency

The number of tallies before each subject given the number of students who like that particular subject.

This is known as the frequency of that subject.

Frequency given the number of times that a particular entry occurs.

From table 5.1, Frequency of students who like English is 4

Frequency of students who like Mathematics is 5

The table made is known as frequency distribution table as it gives the number of times an entry occurs.

Try These

1. A group of students were asked to say which animal they would like most to have as a pet. The results are given below:
   Dog, cat, fat, fish, cat, rabbit, dag, cat, rabbit, dag, cat, dog, dog, dog, cat, cow, fish, rabbit, dog, cat, dog, cat, cat, dog, rabbit, cat, fish, dog.
   Make a frequency distribution table for the same.

Grouping Data

The data regarding choice of subjects showed the occurrence of each of the entries several times. For example, Art is liked by 7 students; Mathematics is liked by 5 students and so in (Above Table). This information can be displayed graphically using a pictograph or a bar graph. Sometimes, however, we have to deal with a large data. For example, consider the following mark (out of 50) obtained in Mathematics by 60 students of a class.

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

If we make a frequency distribution table for each observation, then the table would be too long, so, convenience, we make groups of observations say, 0-10, 10-20 and so on, and obtain frequency distribution of the number of observation falling in each group. Thus, the frequency distribution table to the above data can be.

Groups	**Tally Marks**	**Frequency**
0-10	II	2
10-20	$(IIIII)(IIIII)$	10
20-30	$(IIIII)(IIIII)(IIIII)(IIIII)\ I$	21
30-40	$(IIIII)(IIIII)(IIIII)\ IIII$	19
40-50	$(IIIII)\ II$	7
50-60	I	1
	Total	**60**

Table 3.2

Data presented in this manner is said to be grouped and the distribution obtained is called grouped frequency distribution. It helps us draw meaningful inferences like.

(1) Most of the students have scored between 20 and 40.
(2) Eight students have scored more than 40 marks out of 50 and so on.

Each of the groups 0-10, 10-20, 20-30, etc., is called a class interval (or briefly a class). Observe that 10 occurs in both the classes, i.e., 0-10 as well as 10-20. Similarly, 20 occurs in classes 10-20 and 20-30. But it is not possible that an observation (say 10 of 20) can belong simultaneously to two classes. To avoid this, we adopt the convention that the common observation will belong to the higher class, i.e., 10 belongs of the class interval 10-20 (and not to 0-10). Similarly, 20 belongs to 20-30 (and not to 10-20). In the class interval, 10-20, 10 is called the lower class limit and 20 called the upper class limit. Similarly, in the class interval 20-30, 20 is **lower class limit** and 30 is the **upper class limit.** Observe that the difference between the upper class limit and lower class limit for each of the class intervals 0-10, 10-20, 20-30 etc., is equal, (10 in this class). This difference between the upper class limit and lower class limit is called the width or size of the class interval.

Try these

1. Study the following frequency distribution table and answer the question given below.
    a. What is the size of the class intervals?
    b. Which class has the highest frequency?
    c. Which class has the lowest frequency?
    d. What is the upper limit of the class interval 250-275?
    e. Which two classes have the same frequency?

Class interval (daily income in Afs)	Frequency (Number of workers)
100-125	45
125-150	25
150-175	55
175-200	125
200-225	140
225-250	55
250-275	35
275-300	50
300-325	20
Total	**550**

Table 3.3

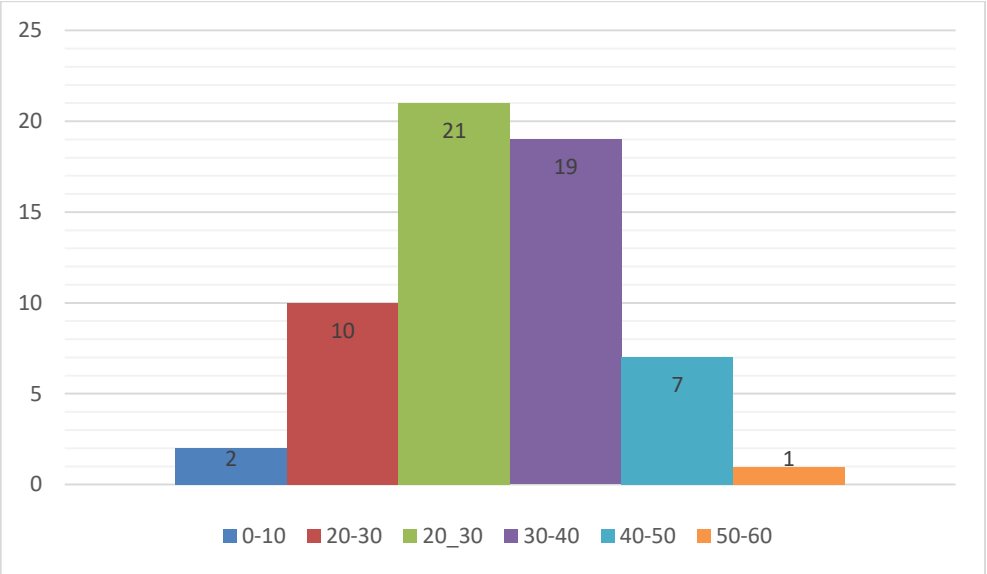
2. Construct a frequency distribution table for the date on weights (in kg) of 20 students of a class using intervals 30-35, 35-40 and so on.
   40, 38, 33, 48, 60, 53, 31, 46, 34, 36, 49, 41, 55, 49, 65, 42, 44, 47, 38, 39.

Bars with a difference: Let us again consider the grouped frequency distribution of the marks obtained by 60 students in Mathematics test.

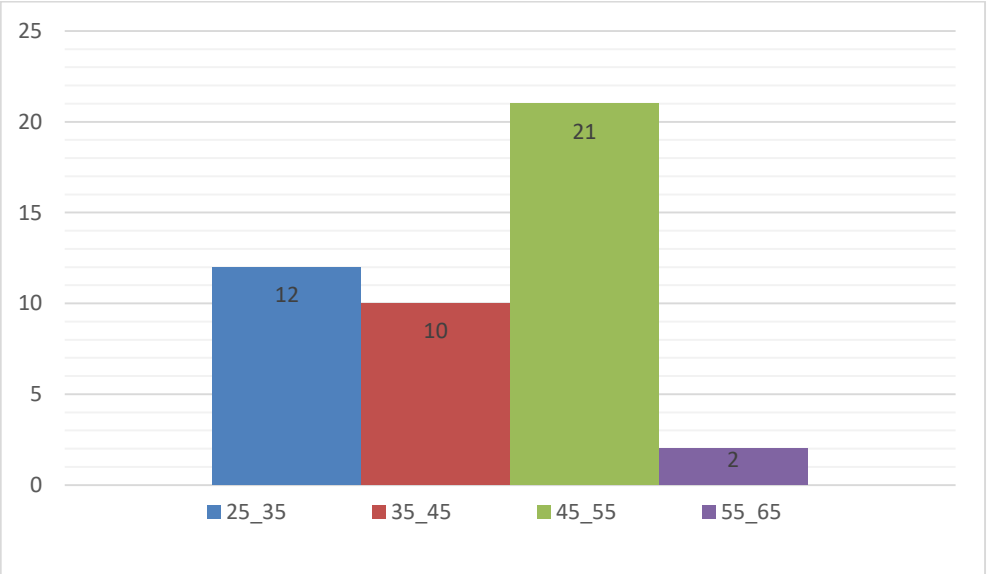
Class Interval	Frequency
0-10	2
10-20	10
20-30	21
30-40	19
40-50	7
50-60	1
Total	**60**

Table 3.4

This is displayed graphically as in the adjoining graph (Fig 3.1).
Is this graph in any way different from the bar graphs which you have drawn in chapter two? Observe that, here we have represented the groups of observation (i.e., class intervals) on the horizontal axis.



The height of the bars shows the frequency of the class-interval. Also, there is on gap between the bars as there is no gap between the class-interval. The graphical representation of data in this manner is called a histogram.
The following graph is another histogram (Fig 3.2).



From the bars of this histogram, we can answer the following questions:

(i) How many teachers are of age 45 years or more but less than 50 years?

(ii) How many teachers are of age less than 35 years?

Try these

1. Observe the histogram (Fig 3.3) and answer the question given below.

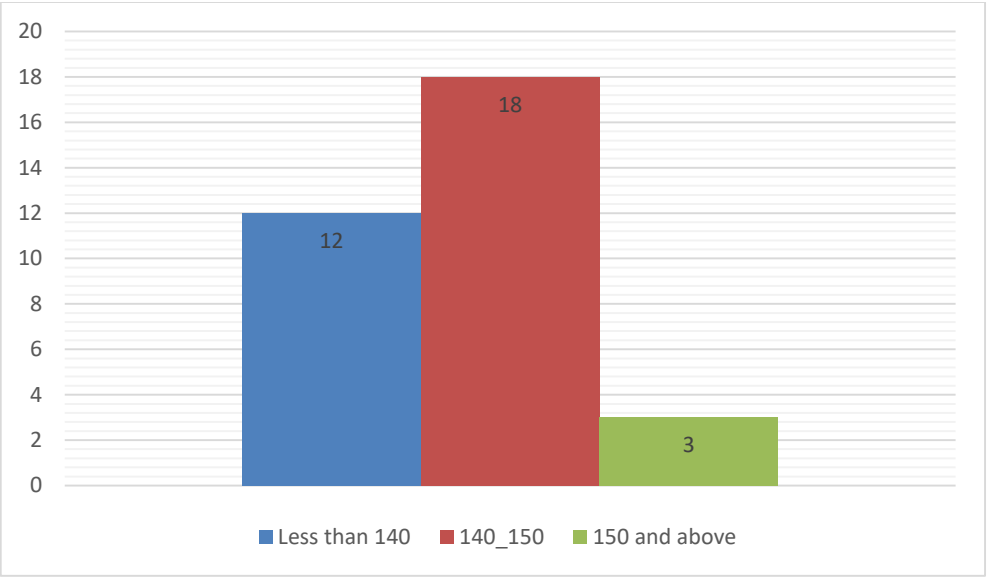


Fig 3.3

(i) What information is being given by the histogram?

(ii) Which group contains maximum girls?

(iii) How many girls have a height of 145 cms and more?

(iv) If we divide the girls into the following three categories, how many would there be in each?

150 cm and more	__Group A
140 cm to less than 150 cm	__Group B
Less than 140 cm	__Group C

Exercise 3.1

1. For which of these would you use a histogram to show the data?

   (a) The number of letters for different areas in a postman's bag.
   (b) The height of competitors in an athletics meet.
   (c) The number of cassettes produced by 5 companies.
   (d) The number of passengers boarding trains from 7:00 am to 7:00 pm. At a station.

Give reasons for each.

2. The shoppers who come to a departmental store are marked as: man (M), woman (W), boy (B), or girl (G). The following list gives the shoppers who came during the first hour in the morning:
   W W W G B W W M G G M M W W W W G B M W B G G M W B G G M W W M M
   W W W M W B W G M W W W W G W M M W W M W G W M G W M M B G G W

   Make frequency distribution table using tally marks. Draw a bar graph to illustrate it.
3. The weekly wages (in Afs) of workers in a factory are.

830, 835, 890, 810, 835, 836, 869, 845, 898, 890, 820, 860, 832, 833, 855,

845, 804, 808, 812, 840, 885, 835, 835, 836, 878, 840, 868, 890, 806, 840,

Using tally marks make a frequency table with intervals as 800-810, 810-820 and so on.

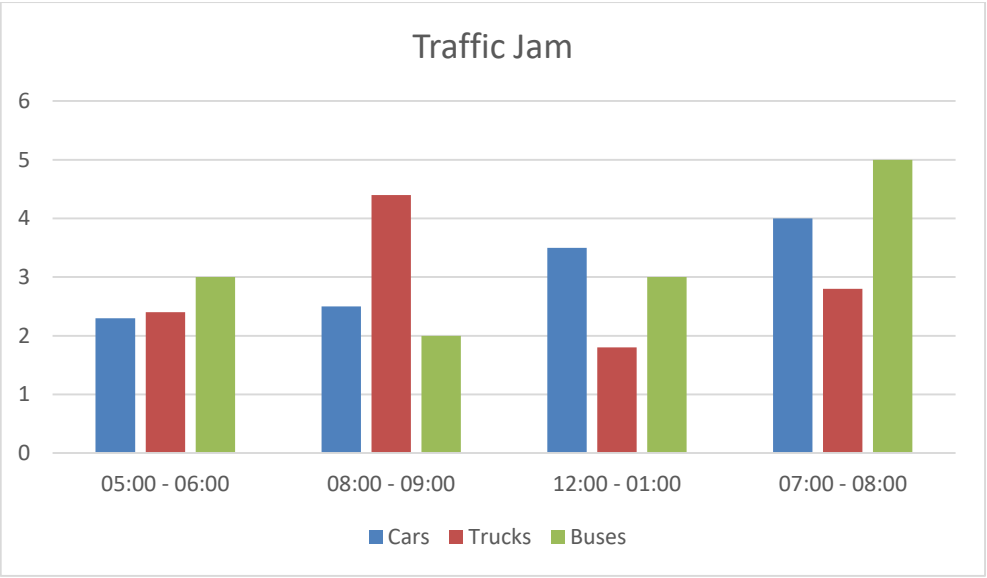
4. Draw a histogram for the frequency table made for the data in question 3, and answer the following question.
   (i) Which group has the maximum number of workers?
   (ii) How many workers earn Afs 850 and more?
   (iii) How many workers earn less than Afs 850?

A bar graph

Representing data by pictograph is not only time consuming but at times difficult too. Let us see some other way of representing data visually. Bars of uniform width can be drawn horizontally or vertically with equal spacing between them and then the length of each bar represents the given number. Such method of representing data is called a bar diagram or a bar graph.

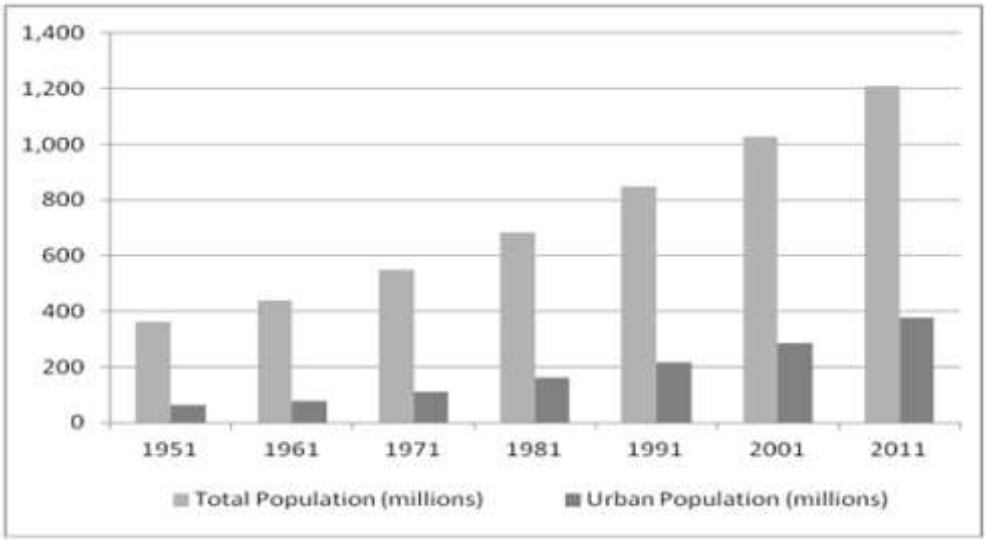
Introduction of a bar graph

Let's look at the example of vehicular traffic at a busy road crossing in Kabul, which was studied by the traffic police on a particular day. The number of vehicles passing through the crossing every hour from 06:00 am to 12:00 noon is shown in the bar graph. One unit of length stands for 100 vehicles.



We can see that maximum traffic is shown by the longest bar (i.e. 5 million vehicles) for the time interval 07:00 – 08:00 A. M. The second longer bar is for 08:00 – 09:00 A. M. Similarly, minimum traffic is shown by the smallest bar for the time interval 05:00 – 06:00 A. M. The bar just longer than the smallest bar is between 12:00 – 01:00 noons.

If the numbers in the data are large, then you may need a different scale. For example, take the case of the growth of the population of India. The numbers are in 10 million. So, if you take 1 unit length to be one person, drawing the bars will not be possible. Therefore, choose the scale as 1 unit to represents 1 million. The bar graph for this case is shown in the figure.



So, the bar of length 200 units represents 200 million and of 8 units represents 800 million.

Example 9: Read the adjoining bar graph showing the number of students in a particular class of a school.

Answer the following questions:

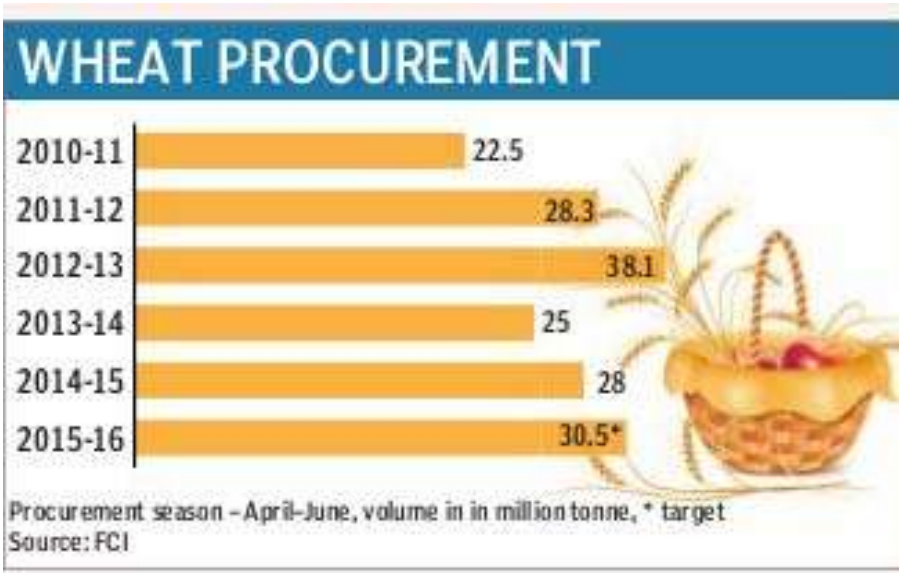
a. What is the scale of this graph?
b. How many new students are added every year?
c. Is the number of students in the year 2003 twice that in the year 2000?

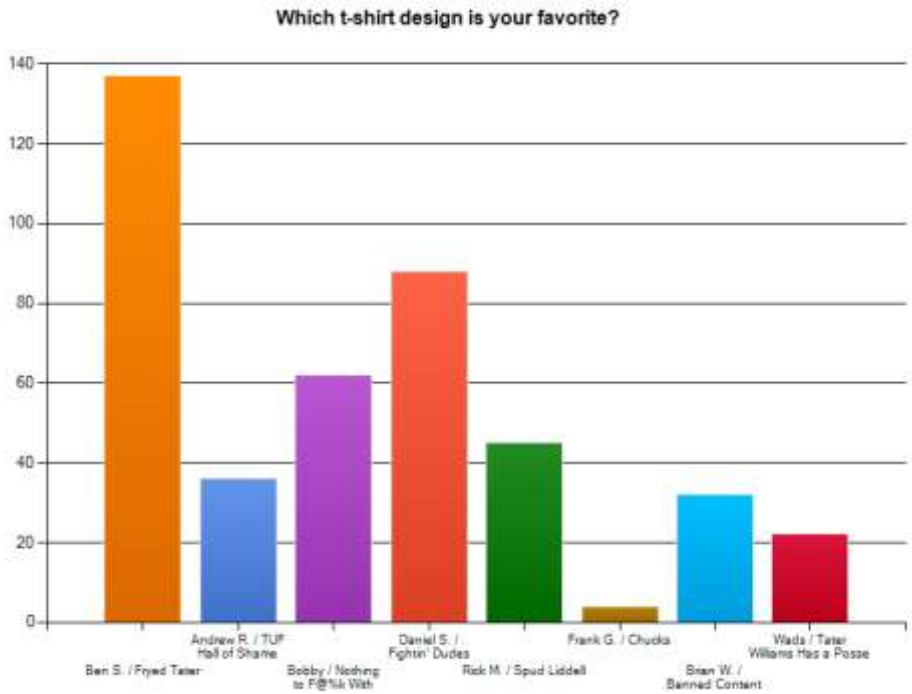
Solution:

a. The scale is 1-unit length equals 10 students. Try (b) and (c) for yourself.

EXERCISE

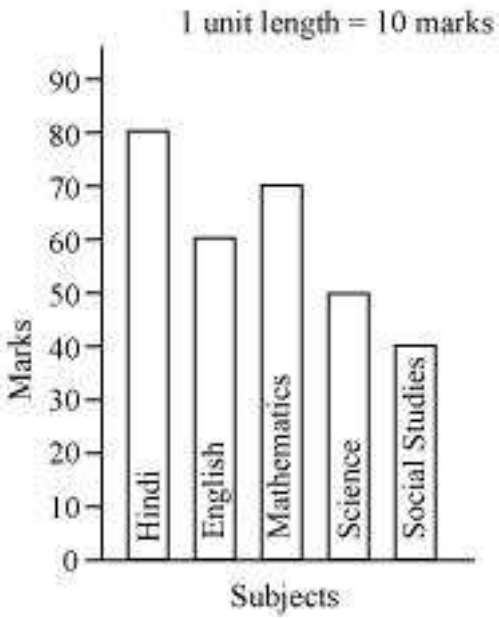
1. The bar graph given alongside shows the amount of wheat purchased by government during the year 1998-2002. Read the bar graph and write down your observations. In which year was
   a. The wheat production maximum?
   b. The wheat production minimum?





2. Observe this bar graph which is showing the scale of shirts in a readymade shop from Monday to Saturday.
   a. What information does the above bar graph give?
   b. What is the scale chosen on the horizontal line representing number of shirts?
   c. On which day were the maximum number of shirts sold? How many shirts were sold on that day?
   d. On which day were the minimum number of shirts sold?
   e. How many shirts were sold on Thursday?

3. Observe this bar graph which shows the marks obtained by Aziz in half-yearly examination in different subjects. Answer the given questions.
   a. What information does the bar graph give?
   b. Name the subject in which Aziz scored maximum marks.
   c. Name the subject in which he has scored minimum marks.
   d. State the name of the subjects and marks obtained in each of them.

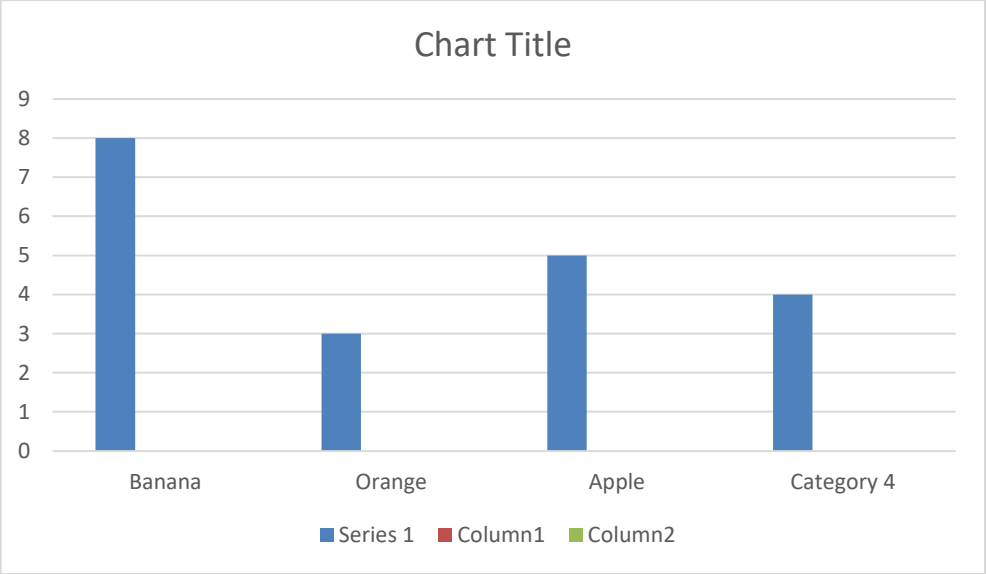


Drawing a bar graph

Recall the example where Aslam had prepared a table representing choices of fruits made by his classmates. Let us draw a bar graph for this data.

Name of fruits	Banana	Orange	Apple	Melon
Number of students	8	3	5	4

First of all draw a horizontal line and a vertical line. On the horizontal line and a vertical line. On the horizontal line we will draw bars representing each fruit and on vertical line we will write numerals representing number of students.
Let us choose a scale. It means we first decide how many students will be represented by unit length of a bar.
Here, we take 1 unit length to represent 1 student only.
We get a bar graph as shown in adjoining figure



Example 10: Following table shows the monthly expenditure of Khalid's family on various items.

To represent this data in the form of a bar diagram, here are the steps.

Items	Expenditure (in Afs)
House rent	3000
Food	3400
Education	800
Electricity	400
Transport	600
Miscellaneous	1200

a. Draw two perpendicular lines, one vertical and one horizontal.
b. Along the horizontal line, mark the "items" and along the vertical line, mark the corresponding expenditure.
c. Take bars of same width keeping uniform gap between them.
d. Choose suitable scale along the vertical line. Let 1 unit length=200Afs and then mark the corresponding values.

Example 11: Calculate the heights of the bars for various items as shown below.

House rent : $3000 \div 200 = 15\ units$
Food : $3400 \div 200 = 17\ units$
Education : $800 \div 200 = 4\ units$
Electricity : $400 \div 200 = 2\ units$
Transport : $600 \div 200 = 3\ units$
Miscellaneous : $1200 \div 200 = 6\ units$

DO THIS

1. Along with your friends, think of five more situations where we can have data. For this data, construct the tables and represent them using bar graphs.

EXERCISE

1. A survey of 120 school students was done to find which activity they prefer to do in their free time.

Preferred activity	Number of students
Playing	45
Reading story books	30
Watching TV	20
Listening to music	10
Painting	15

Draw a bar graph to illustrate the above data taking scale of 1 unit length = 5 students.
Which activity is preferred by most of the students other than playing?

2. The number of Mathematics books sold by a shopkeeper on six consecutive days is shown below:

Days	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
Number of books sold	65	40	30	50	20	70

Draw a bar graph to represent the above information choosing the scale of your choice.

3. Following table shows the number of bicycles manufactured in a factory during the years 1998 to 2002. Illustrate this data using a bar graph. Choose a scale of your choice.

Years	**Number of bicycles manufactured**
1998	800
1999	600
2000	900
2001	1100
2002	1200

a. In which year were the maximum number of bicycles manufactured?
b. In which year were the minimum number of bicycles manufactured?

4. Number of persons in various age group in a town is given in the following table.

Age group	1-14	15-29	30-44	45-59	60-74	75 and above
Number of persons	200000	160000	120000	120000	80000	40000

Draw a bar graph to represent the above information and answer the following questions. (Take 1-unit length = 20 thousands)

a. Which two age groups have same population?
b. All persons in the age group of 60 and above are called senior citizens. How many senior citizens are there in the town?

Use of bar graphs with a different purpose

We have seen in chapter one how information collected could be first arranged in a frequency distribution table and then this information could be put as a visual representation in the form of a pictographs or bar graphs.

You can look at the bar graphs and make deductions about the data. You can also get information based on these bar graphs. For example, you can say that the mode is the longest bar if the bar represents the frequency.

Choosing a scale

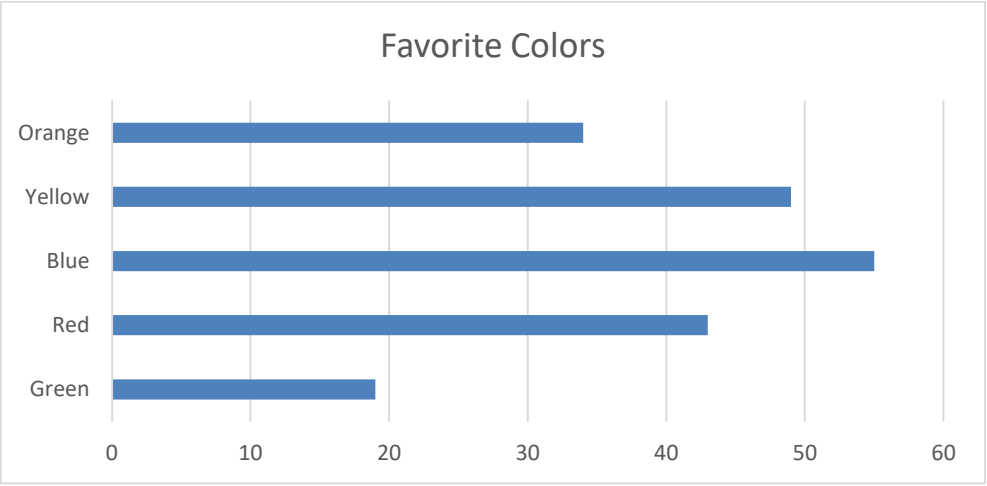
We know that a bar graph is a representation of numbers using bars of uniform width and the lengths of the bars depend upon the frequency and the scale you have chosen. For example, in a bar graph where numbers in units are to be shown, the graph represents one unit length for one observation and if it has to show numbers in tens or hundreds, one unit length can represent 10 or 100 observations. Consider the following example:

EXAMPLE: Two hundred students of 6^{th} and 7^{th} class were asked to name their favorite color so as to decide upon what should be the color of their school building. The results are shown in the following table. Represent the given data on a bar graph.

Favorite color	Red	Green	Blue	Yellow	Orange
Number of students	43	19	55	49	34

Answer the following questions with the help of the bar graph:

a. Which is the most preferred color and which is the least preferred?
b. How many colors are there in all? What are they?



Solution

Choose a suitable scale as follows:
Start the scale at 0. The greatest value in the data is 55, so end the scale at a value greater than 55, such as 60. Use equal divisions along the axes, such as increments of 10. You know that all the bars would lie between 0 and 60. We choose the scale such that the length between 0 and 60 is neither too long nor too small. Here we take 1 unite for 10 students. We then draw and label the graph as shown. From the bar graph we conclude that

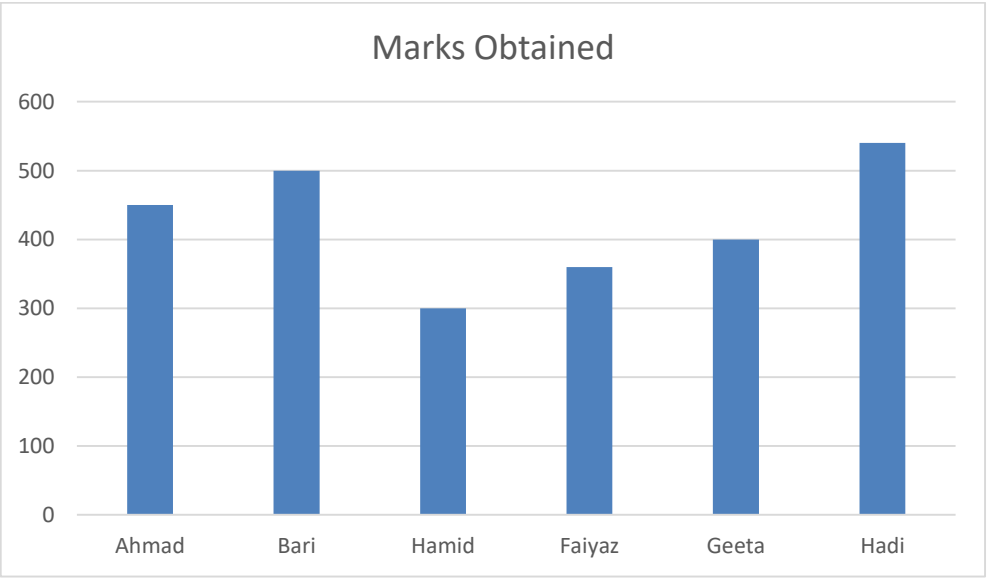
a. Blue is the most preferred color (because the bar representing blue is the tallest).
b. Green is the least preferred color. (Because the bar representing green is the shortest).
c. There are five colors. They are red, green, blue, yellow and orange. (these are observed on the horizontal line)

EXAMPLE: Following data gives total marks (out of 600) obtained by six children of a particular class. Represent the data on a bar graph.

Students	Ahmad	Bari	Hamid	Faiyaz	Geeta	Hadi
Marks Obtained	450	500	300	360	400	540

Solution:

a. To choose an appropriate scale we make equal divisions taking increments of 100. Thus 1 unit will represent 100 marks. (What would be the difficulty if we choose one unit to represent 10 marks?)
b. Now represent the data on the data on the bar graph.



Drawing double bar graph

Consider the following two collections data giving the average daily hours of sunshine in two cities Kabul and Jalalabad for all the twelve months of the year. These cities are near the South Pole and hence have only a few hours of sunshine each day.

In Jalalabad												
	Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct	Nov	Dec
Average hours of sunshine	2	$3\frac{1}{4}$	4	4	$7\frac{3}{4}$	8	$7\frac{1}{2}$	7	$6\frac{1}{4}$	6	4	2
In Kabul												
Average hours of sunshine	$1\frac{1}{2}$	3	$3\frac{1}{2}$	6	$5\frac{1}{2}$	$6\frac{1}{2}$	$5\frac{1}{2}$	5	$4\frac{1}{2}$	4	3	$1\frac{3}{4}$

By drawing individual bar graphs you could answer questions like

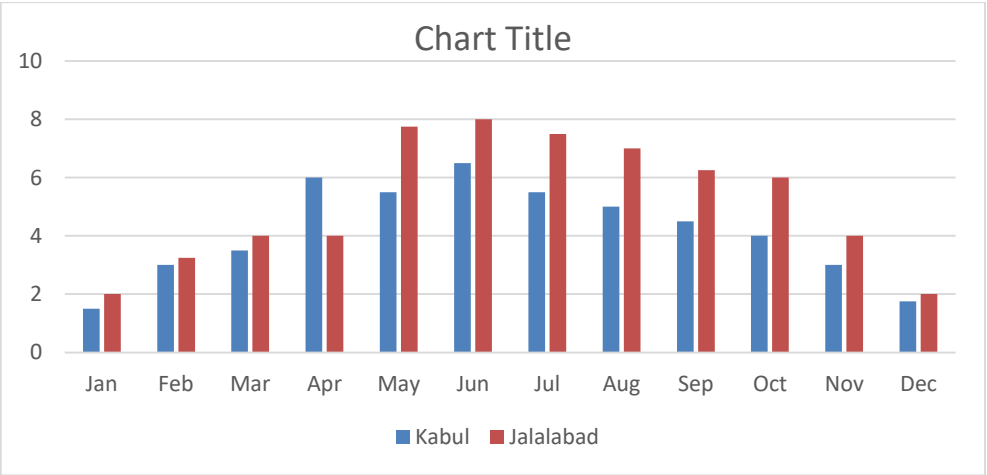
a. In which month does each city has maximum sunlight?

b. In which months does each city has minimum sunlight?

However, to answer questions like “In a particular month, which city has more sunshine hours”, we need to compare the average hours of sunshine of both the cities. To do this we will learn to draw what is called a double bar graph giving the information of both cities side by side.

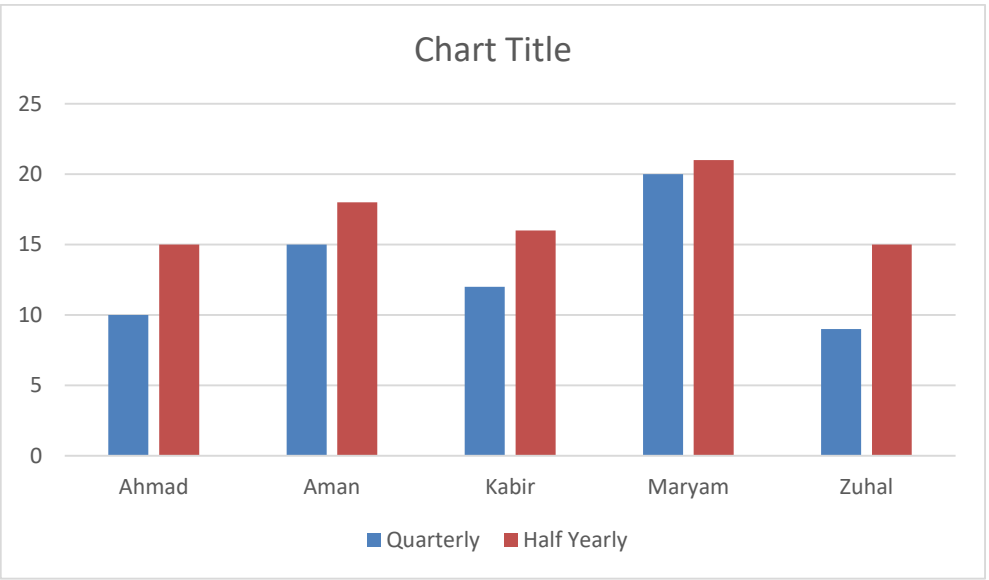
The following bar graph shows the average sunshine of both the cities.

For each month we have two bars, the heights of which give the average hours of sunshine in each city. From this we can infer that except for the month of April, there is always more sunshine in Jalalabad than in Kabul. You could put together a similar bar graph for your area or for your city.



EXAMPLE: A mathematics teacher wants to see, whether the new technique of teaching he applied after quarterly test was effective or not. He takes the scores of the 5 weaker students in the quarterly test (out of 25) and in the half yearly test (out of 25).

Students	Ahmad	Aman	Kabir	Maryam	Rita
Quarterly	10	15	12	20	9
Half yearly	15	18	16	21	15



Solution: He draws the adjoining double bar graph and finds a marked improvement in most of the students, the teacher decides that he should continue to use the new technique of teaching. Can you think of a few more situations where you could use double bar graph?

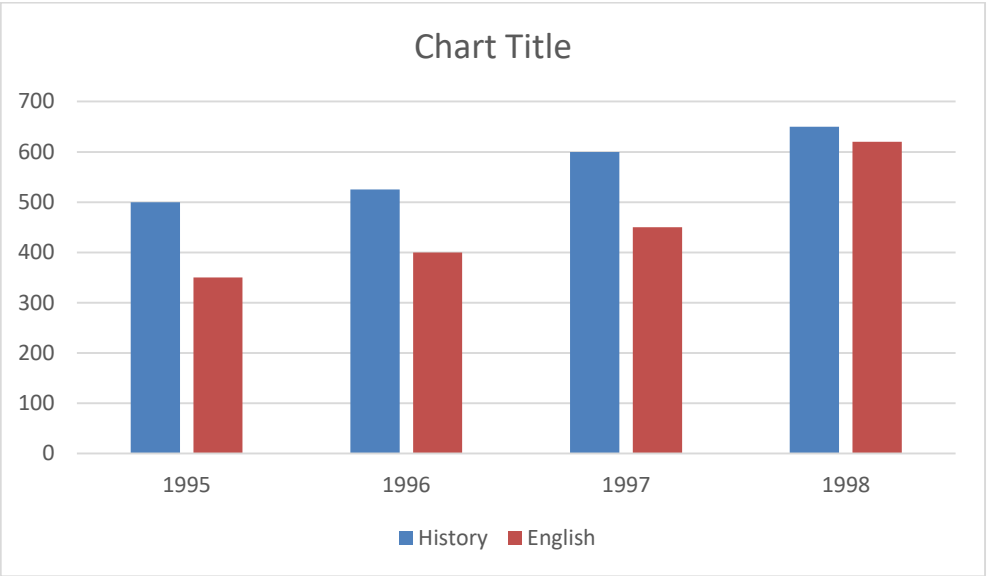
TRY THESE

1. Sale of English and history books in the year 1995, 1996, 1997 and 1998 are given below:

Years	1995	1996	1997	1998
English	350	400	450	620
History	500	525	600	650

Draw a double bar graph and answer the following questions:

a. In which year was the difference in the sale of the two field books least?
b. Can you say that the demand for English books rose faster? Justify.



EXERCISE

1. Read the above bar graph which shows the number of books sold by a bookstore during five consecutive years and answer the following questions:

a. About how many books were sold in 1989, 1990, 1992?
b. In which year were about 475 books sold? About 225 books sold?
c. In which years were fewer than 250 books sold?
d. Can you explain how you would estimate the number of books sold in 1989?

2. Number of children in six different classes are given below. Represent the data on a bar graph?

Class	Fifth	Sixth	Seventh	Eighth	Ninth	Tenth
Number of children	135	120	95	100	90	80

a. How would you choose a scale?
b. Which class has the maximum number of children? And the minimum?
c. Find the ratio of students of class sixth to the students of class eight.

3. The performance of a student in 1st term and 2nd term is given. Draw a double bar graph choosing appropriate scale and answer the following:

Subject	English	History	Math	Science	Accounting
1st Term	67	72	88	81	73
2nd Term	70	65	95	85	75

a. In which subject, has the child improved his performance the most?
b. In which subject is the improvement the least?
c. Has the performance gone down in any subject?

4. Consider this data collected from a survey of a colony.

Favorite Sport	cricket	Basket Ball	Swimming	Hockey	Athletics
Watching	1240	470	510	430	250
participating	620	320	320	250	105

a. Draw a double bar graph choosing an appropriate scale. What do you infer from the bar graph?
b. Which sport is most popular?
c. Which is more preferred, watching or participating in sports?

5. Take the data giving the minimum and the maximum temperature of various cities given in the beginning of this chapter. Plot a double bar graph using the data and answer the following:

a. Which city has the largest difference in the minimum and maximum temperature on the given date?
b. Which is the hottest city and which is coldest city?
c. Name two cities where maximum temperature of one was less than the minimum temperature of the other.
d. Name the city which has the least difference between its minimum and the maximum temperature.

TRY THESE: Think of some situations, at least 3 examples of each, that are certain to happen, some that are impossible and some that may or may not happen i. e., situations that have some chance of happening.

Histogram

This is a form of representation like the bar graph, but it is used for continuous class intervals. For instance, consider the frequency distribution table 4.6, representing the weights of 36 students of a class:

Table 4.6

Weights (in kg)	Number of Students
$30.5 - 35.5$	9
$35.5 - 40.5$	6
$40.5 - 45.5$	15
$45.5 - 50.5$	3
$50.5 - 55.5$	1
$55.5 - 60.5$	2
Total	**36**

Let us represent the data given above graphically as follows:

a. We represent the weights on the horizontal axis on a suitable scale. We can choose the scale as 1 cm = 5 kg. Also, since the first class interval is starting from 30.5 and not zero, we show it on the graph by making a kink or a break on the axis.
b. We represent the number of students (frequency) on the vertical axis on a suitable scale. Since the maximum frequency is 15, we need to choose the scale to accommodate this maximum frequency.
c. We now draw rectangles (or rectangular bars) of width equal to the class size and lengths according to the frequencies of the corresponding class intervals. For example, the rectangle for the class interval 30.5-35.5 will be of width 1 cm and length 4.5 cm.
d. In this way, we obtain the graph as shown in fig 4.3.

Example 4. Let us now consider the following frequency distribution table which gives the weights of 38 students of a class:

Weights (in Kg)	Number of students
31-35	9
36-40	5
41-45	14
46-50	3
51-55	1
56-60	2
61-65	2
66-70	1
71-75	1
Total	**38**

Now, if two new students of weights 35.5 kg and 40.5 kg are admitted in this class, then in which interval will we include them? We cannot add them in the ones ending with 35 or 40, nor to the following ones. This is because there are gaps in between the upper and lower limits of two consecutive classes. So, we need to divide the intervals so that the upper and lower limits of consecutive intervals are the same. For this, we find the difference between the upper limit of a class and the lower limit of its succeeding class. We then add half of this difference to each of the upper limits and subtract the same from each of the lower limits.
For example, consider the classes $31 - 35$ and $36 - 40$.
The lower limit of $36 - 40 = 36$.
The upper limit of $31 - 35 = 35$.
The difference $= 36 - 35 = 1$.
So, half of the difference $= \frac{1}{2} = 0.5$.
So the new class interval formed from $31 - 35$ is $(31 - 0.5) - (35 + 0.5)$, i.e., $30.5 - 35.5$. Similarly, the new class formed from the class $36 - 40$ is $(36 - 05) - (40 + 0.5)$, i. e., $35.5 - 40.5$.
Observe that since there are no gaps in between consecutive rectangles, the resultant graph appears like a solid figure. This is called a histogram, which is a graphical representation of a grouped frequency distribution with continuous classes. Also, unlike a bar graph, the width of the bar plays a significant role in it construction.
Here, in fact, areas of the rectangles erected are proportional to the corresponding frequencies. However, since the width of the rectangles are all equal, the lengths of the rectangles are proportional to the frequencies. That is why, we draw the lengths according to (c) above.
Now, consider a situation different from the one above.
Example 7: A teacher wanted to analyze the performance of two sections of students in a mathematics test of 100 marks. Looking at their performances, she found that a few students got under 20 marks and a few got 70 marks or above. So she decided to group them into intervals of varying sizes as follows: 0-20, 20-30,… 60-70, 70-100. Then she formed the following table:

Table 4.7

Marks	**Number of students**
0-20	7
20-30	10
30-40	10
40-50	20
50-60	20
60-70	15
70-100	8
Total	**90**

Carefully examine this graphical representation. Do you think that it correctly represents the data? No, the graph is giving us a misleading picture. As we have mentioned earlier, the areas of the rectangles are proportional to the frequencies in a histogram. Earlier this problem did not arise, because the widths of all the rectangles were equal. But here, since the widths of the rectangles are varying, the histogram above does not give a correct picture. For example, it shows a greater frequency in the interval 70-100, than in 60-70, which is not the case.

So, we need to make certain modifications in the lengths of the rectangles so that the areas are again proportional to the frequencies.

The steps to be followed are as given below:

1. Select a class interval with the minimum class size. In the example above, the minimum class size is 10.
2. The lengths of the rectangles are then modified to be proportionate to the class size 10.

For instance, when the class size is 20, the length of the rectangle is 7. So when the class size is 10, the length of the rectangle will be $\frac{7}{20} \cdot 10 = 3.5$.

Similarly, proceeding in this manner, we get the following table:

Table 4.8.

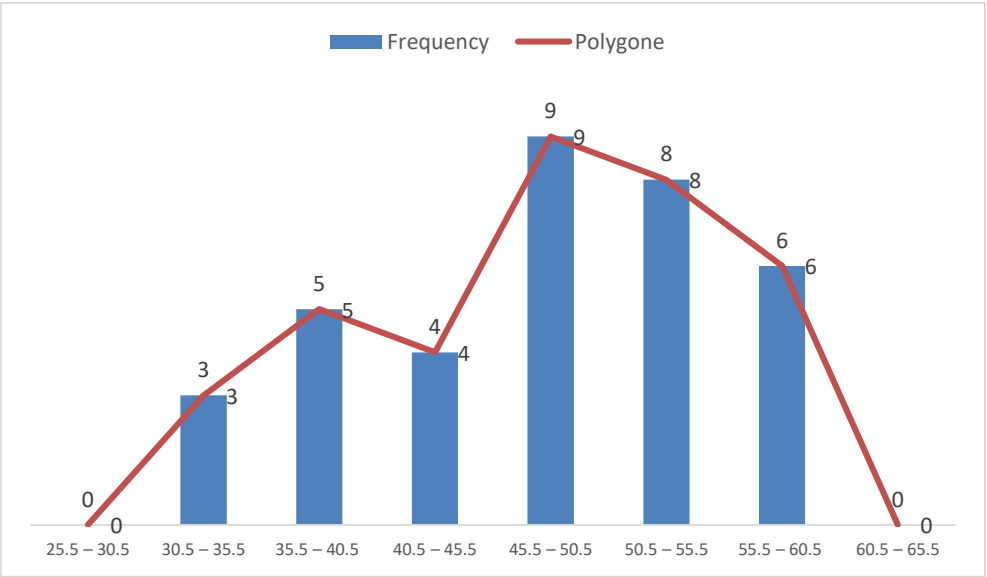
Marks	Frequency	Width of the class	Length of the rectangle
0-20	7	20	$\frac{7}{20} \cdot 10 = 3.5$
20-30	10	10	$\frac{10}{10} \cdot 10 = 10$
30-40	10	10	$\frac{10}{10} \cdot 10 = 10$
40-50	20	10	$\frac{20}{10} \cdot 10 = 20$
50-60	20	10	$\frac{20}{10} \cdot 10 = 20$
60-70	15	10	$\frac{15}{10} \cdot 10 = 15$
70-100	8	30	$\frac{8}{30} \cdot 10 = 2.67$

Since we have calculated these lengths for an interval of 10 marks in each case, we may call these lengths as “proportion of students per 10 marks interval”. So, the correct histogram with varying width is given in fig. 4.5.

Frequency polygon:

There is yet another visual way of representing Quantitative data and its frequencies. This is a polygon. To see what we mean, consider the histogram represented by fig 4.3. Let us call these mid points B, C, D, E, F and G. when joined by line segments, we obtain the figure BCDEFG (see fig 4.6). to complete the polygon, we assume that there is a class interval with frequency zero before 30.5-35.5, and one after 55.5-60.5, and their mid points are A and H, respectively. ABCDEFGH is the frequency polygon corresponding to the data shown in fig 4.3. we have shown this in fig 4.6.

Marks	Frequency
25.5 – 30.5	0
30.5 – 35.5	3
35.5 – 40.5	5
40.5 – 45.5	4
45.5 – 50.5	9
50.5 – 55.5	8
55.5 – 60.5	6
60.5 – 65.5	0
Total	**35**



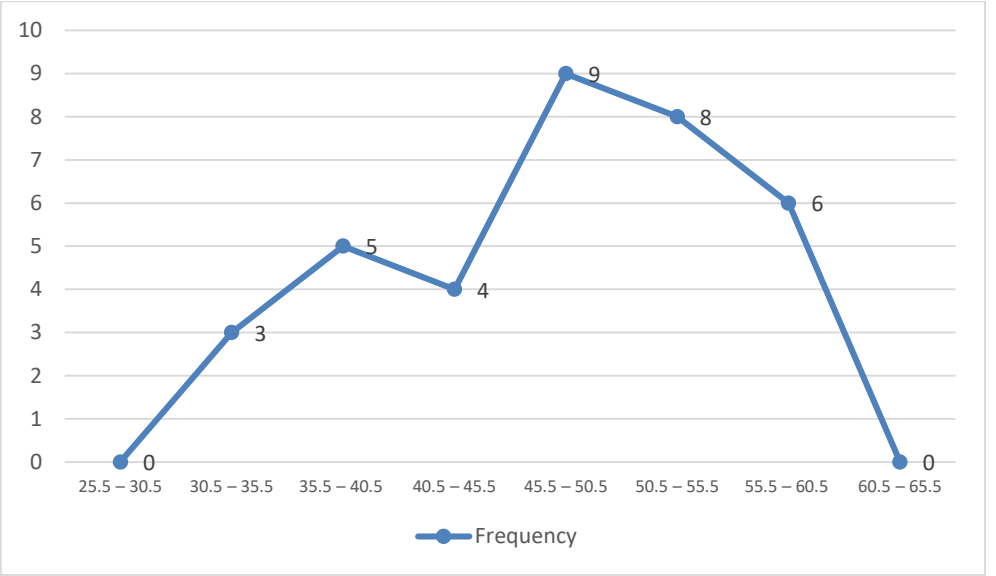


Fig 4.6.

Although, there exists no class preceding the lowest class and no class succeeding the highest class, addition of the two class intervals with zero frequency enables us to make the area of the frequency polygon the same as the area of the histogram. Why is this so? Now, the question arises: how do we complete the polygon when there is no class preceding the first class? Let us consider such a situation.

Example 8: consider the marks, out of 100, obtained by 51 students of a class in a test, given in table 4.9.

Marks	**Number of students**
$0-10$	5
$10-20$	10
$20-30$	4
$30-40$	6
$40-50$	7
$50-60$	3
$60-70$	2
$70-80$	2
$80-90$	3
$90-100$	9
Total	**51**

Draw a frequency polygon corresponding to this frequency distribution table.

Solution: Let us first draw a histogram for this data and mark the mid points of the tops of the rectangles as B, C, D, E, F, G, H, I, J, K, respectively. Here, the first class is 0-10. So, to find the class preceding 0-10, we extend the horizontal axis in the negative direction and the midpoint of the imaginary class interval (-10)-0. The first end point, i.e., B is joined to this midpoint with zero frequency on the negative direction of the horizontal axis. The point of the class succeeding the last class of the given data. Then OABCDEFGHIJKL is the frequency polygon, which is shown in fig 4.7.

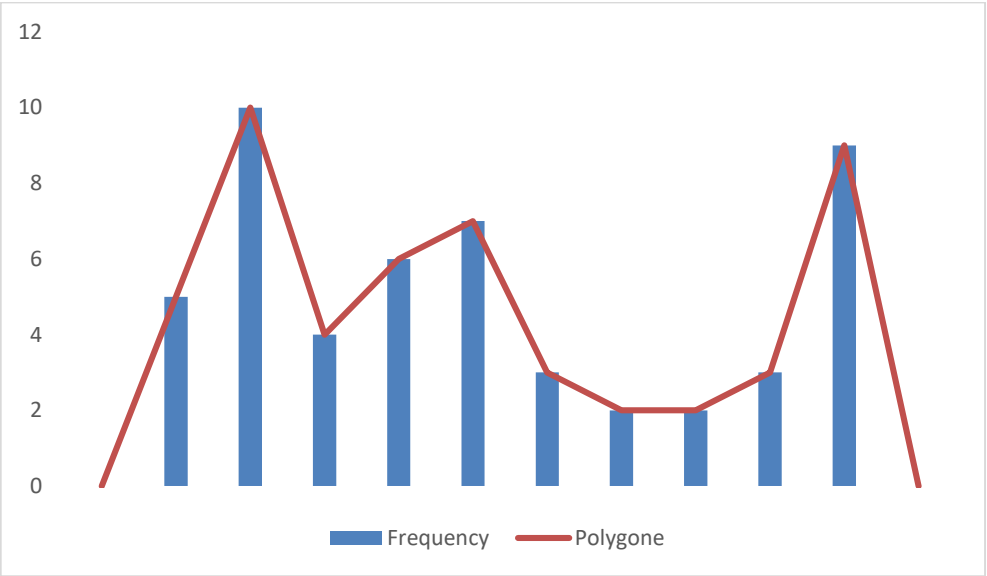


Fig 4.7

Frequency polygons can also be drawn independently without drawing histograms. For this, we require the midpoints of the class intervals used in the data. These midpoints of the class interval are called class marks.

To find the class mark of a class interval, we find the sum of the upper limit and lower limit of a class and divide it by 2. Thus,

$$Class\ Mark = \frac{Upper\ Limit + Lower\ Limit}{2},$$

Example 9: in a city, the weekly observations made in a study on the cost of living index are given in the following table:

Table 4.10

Cost of living index	Number of weeks
140-150	5
150-160	10
160-170	20
170-180	9
180-190	6
190-200	2
Total	**52**

Draw a frequency polygon for the data above (without constructing a histogram).

Solution: Since we want to draw a frequency polygon without a histogram, let us find the class marks of the classes given above, that is of 140-150, 150-160,

So, the class mark$= \frac{150+140}{2} = \frac{290}{2} = 145.$

Continuing in the same manner, we find the class marks of the other classes as well.

So, the new table obtained is as shown in the following table:

Table 4.11

Classes	Class Marks	Frequency
140-150	145	5
150-160	155	10
160-170	165	20
170-180	175	9
180-190	185	6
190-200	195	2
Total		**52**

We can draw a frequency polygon by plotting the class marks along the horizontal axis, the frequencies along the vertical axis, and then plotting and joining the points $B(145,5), C(155,10), D(165,20), E(175,9), F(185,6),$ $,and\ G(195,2)$ by line segments.

We should not forget to plot the point corresponding to the class mark of the class 130-140 (just before the lowest class 140-150) with zero frequency, that is, $A(135,0)$, and the point $H(205,0)$, occurs immediately after $G(195,2)$. So, the resultant frequency polygon will be ABCDEFGH (see fig 4.8).

Frequency polygons are used when the data is continues and very large. It is very useful for comparing two different sets of data of the same nature, for example, comparing the performance of two different sections of the same class.

Chapter Exercise

1. A survey conducted by an organization for the cause of illness and death among the women between the ages 15-44 (in years) worldwide, 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

a. Represent the information given above graphically.
b. Which condition is the major cause of women's ill health and death worldwide?
c. Try to find out, with the help of your teacher, any two factors which play a major role in the cause in (b) above being the major cause.

2. The following data on the number of girls (to the nearest ten) per thousand boys in different sections of Indian 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

a. Represent the information above by a bar graph.
b. In the classroom discuss what conclusions can be arrived at from the graph.

3. Given below are the seats won by different political parties in the polling outcome of a state assembly elections:

Political party	A	B	C	D	E	F
Seats won	75	55	37	29	10	37

a. Draw a bar graph to represent the polling results.

b. Which political party won the maximum number of seats?

4. The length of 40 leaves of a plant are measured correct to one millimeter, and the obtained data is represented in the following table:

Length (in mm)	Number of leaves
118-126	3
127-135	5
136-144	9
145-153	12
154-162	5
163-171	4
172-180	2

a. Draw a histogram to represent the given data. {Hint: first make the class intervals continuous}.
b. Is there any other suitable graphical representation for the same data?
c. Is it correct to conclude that the maximum number of leaves are 153 mm long? Why?

5. The following table gives the life times of 400 neon lamps:

Life time (in hours)	Number of lamps
300-400	14
400-500	56
500-600	60
600-700	86
700-800	74
800-900	62
900-1000	48

a. Represent the given information with the help of a histogram.
b. How many lamps have a life time of more than 700 hours?

6. The following table gives the distribution of students of two sections according to the marks obtained by them:

Section A		Section B	
Marks	Frequency	Marks	Frequency
0-10	3	0-10	5
10-20	9	10-20	19
20-30	17	20-30	15
30-40	12	30-40	10
40-50	9	40-50	1

Represent the marks of the students of both the sections on the same graph by two frequency polygons. From the two polygons compare the performance of the two sections.

7. 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
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. {Hint: First make the class intervals continuous}.

8. A random survey of the number of children of various age groups playing in a park was found as follows:

Age (in years)	Number of children
1-2	5
2-3	3
3-5	6
5-7	12
7-10	9
10-15	10
15-17	4

Draw a histogram to represent the data above.

9. 100 surnames were randomly picked up from a local telephone directory and a frequency distribution of the number of letters in the English alphabet in the surnames was found as follows:

Number of letters	Number of surnames
1-4	6
4-6	30
6-8	44
8-12	16
12-20	4

a. Draw a histogram to depict the given information.
b. Write the class interval in which the maximum number of surnames lie.

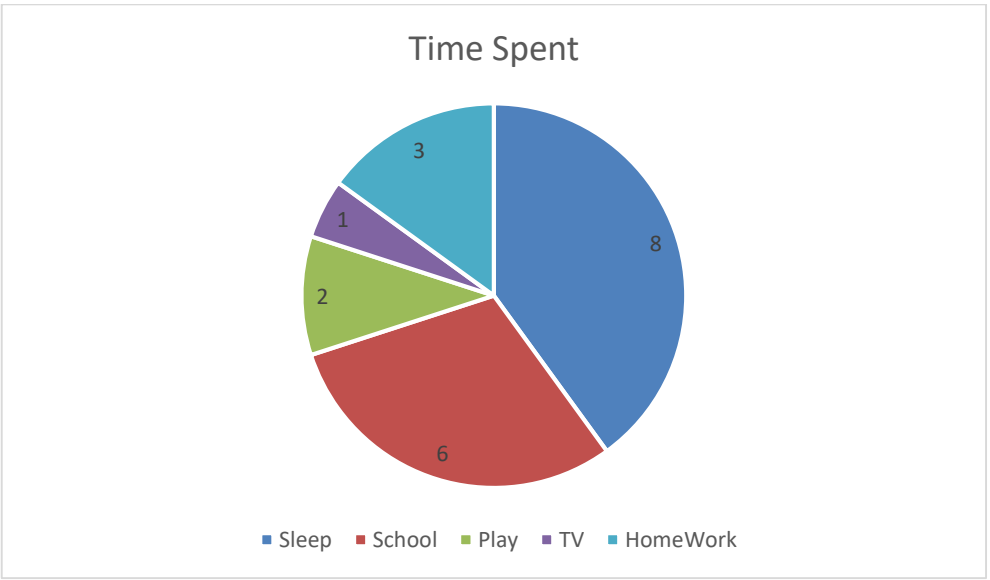
3

One Variable Graphs

Pie Chart

Have you come across data represented in circular form as shown (Fig 5.4)?

The time spent by a child during a day



These are called circle graphs. A circle graph shows the relationship between a whole and its parts. Here, the whole circle is divided into sectors. The size of each sector is proportional to the activity of information it represents.

For example, in the above graph, the proportion of the sector for hours spent in sleeping

$$\frac{number\ of\ sleeping\ hours}{whole\ day} = \frac{8\ hour}{24\ hour} = \frac{1}{3}.$$

So, this sector is drawn as $\frac{1}{3}rd$ part of the circle. Similarly, the proportion of the sector for hours spent in school is given as follows:

$$\frac{number\ of\ school\ hours}{whole\ day} = \frac{6\ hours}{24\ hours} = \frac{1}{4}$$

So this sector is drawn $\frac{1}{4}th$ of the circle. Similarly, the size of other sectors can be found.

Add up the fractions for all the activities. Do you get the total as one?

A circle graph is also called a pie chart.

Try These

1. Each of the following pie charts (Fig 3.5) gives you a different piece of information about your class.
   Find the fraction of the circle representing each of these information.
2. Answer the following questions based on the pie chart given (fig 3.6)

(i) Which type of programs are viewed the most?

(ii) Which two types of programs have number of viewers equal to those watching sports channels?

Drawing pie charts

The favorite flavors of ice-creams for students of a school are given in percentages as follows.

Table 3.5

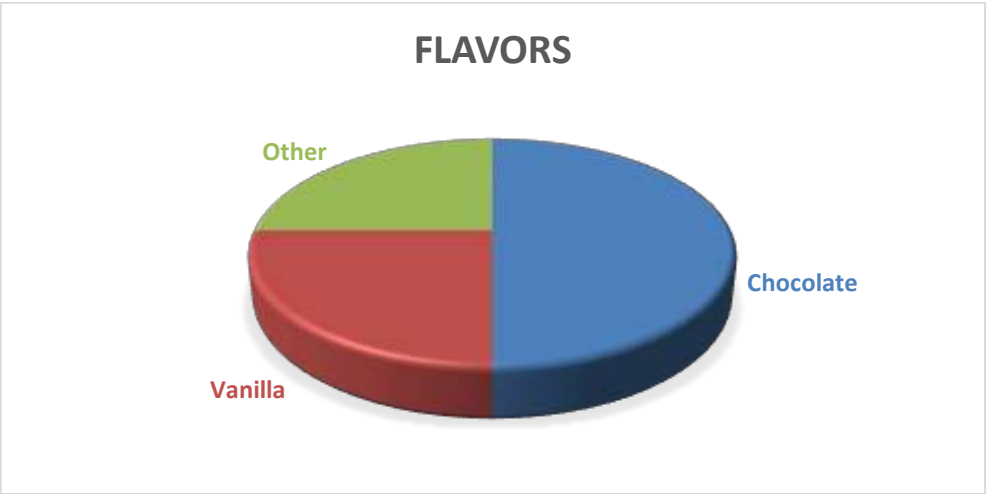
Flavors'	**Percentage of students preferring the flavors**
Chocolate	50%
Vanilla	25%
Other flavors	25%

Let us represent this data in a pie chart.

The total angle at the center of a circle is 360 degrees. The central angle of the sectors will be a fraction of 360. We make a table to find the central angle of the sectors (Table 3.5).

Flavors	**Students in percent preferring the flavors**	**In fractions**	**Fraction of 360^o**
Chocolate	50%	$\frac{50}{100} = \frac{1}{2}$	$\frac{1}{2}\ of\ 360 = 180$
Vanilla	25%	$\frac{25}{100} = \frac{1}{4}$	$\frac{1}{4}\ of\ 360^o = 90^o$
Other Flavors	25%	$\frac{25}{100} = \frac{1}{4}$	$\frac{1}{4}\ of\ 360^o = 90^o$

- Draw a circle with any convenient radius. Mark its center (O) and a radius (OA)
- The angle of the sector for chocolate is 180 degrees. Use the protractor to draw $< AOB = 180^o$
- Continue marking the remaining sectors.



Example 1: adjoining pie chart (Fig 3.7) gives the expenditure (in percentage) on various items and savings of family during a month.

(i) On which item, the expenditure was maximum?

(ii) Expenditure on which item is equal to the total saving of the family?

(iii) If the monthly saving of the family is Afs 3000,
What is the monthly expenditure in clothes?

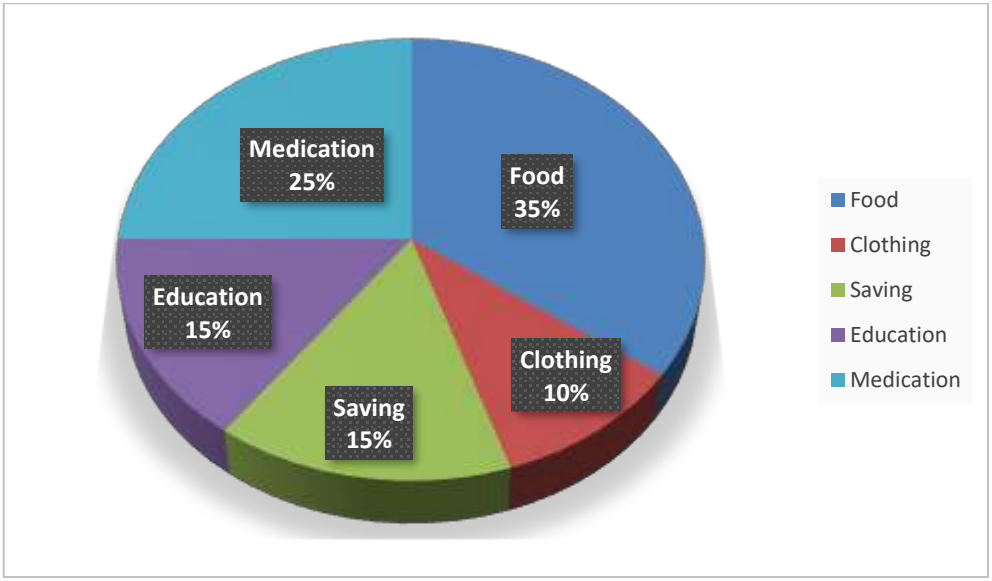


Fig 3.7

Solution:

(i) Expenditure is maximum on food.

(ii) Expenditure on education of children is the same (i.e., 15%) as the saving of the family.

(iii) 15% represents Afs 3000

Therefore, 10 % represents Afs $\frac{3000}{15} \times 10 = Afs\ 2000$.

Example 2: On a particular day, the sales (in Afs) of different items of a baker's shop are given below.

Ordinary bread	320
Biscuits	80
Cakes and pastries	160
Fruit bread	80
Others	40
Total	**720**

Solution: we find the central angle of each sector. Here the total sale =Rs 720. We thus have this table.

Item	Sales (in Afs)	In Fraction	Central Angle
Ordinary bread	320	$\frac{320}{720}=\frac{4}{9}$	$\frac{4}{9}\times 360^o = 160^o$
Biscuits	120	$\frac{120}{720}=\frac{1}{6}$	$\frac{1}{6}\times 360^o = 60^o$
Cakes and pastries	160	$\frac{160}{720}=\frac{2}{9}$	$\frac{2}{9}\times 360^o = 80^o$
Fruit bread	80	$\frac{80}{720}=\frac{1}{9}$	$\frac{1}{9}\times 360^o = 40^o$
others	40	$\frac{40}{720}=\frac{1}{18}$	$\frac{1}{18}\times 360^o = 20^o$

Now, we make the pie chart (fig 3.8)

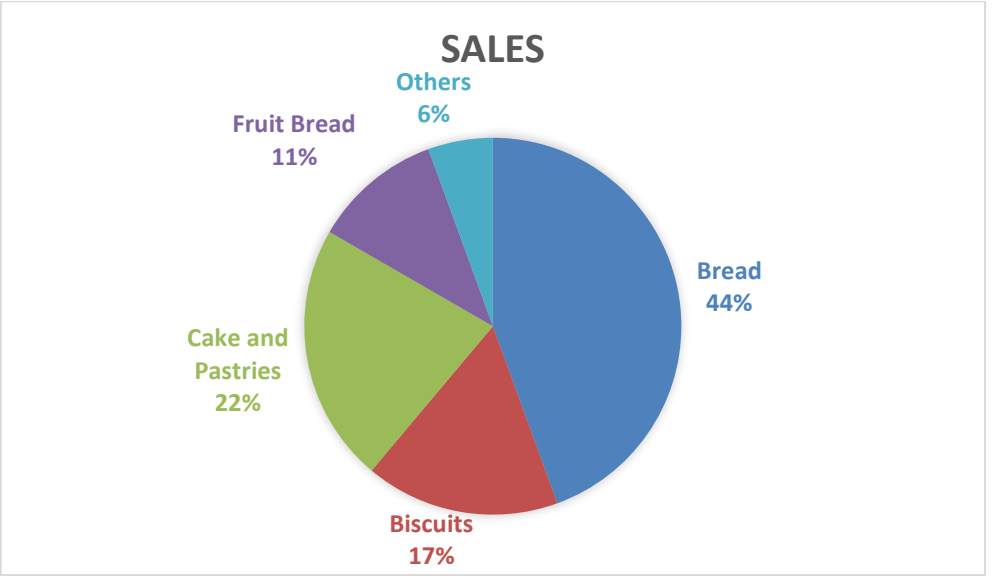


Fig 3.8

Try These

Draw a pie chart of the data given below. The time spent by a child during a day.

Sleep ----------8 hours
School---------- 6 hours
Home work-----4 hours
Play----------4 hours
Other----------2 hours

Think, Discuss and Write

Which form of graph would be appropriate to display the following data?

Production of food grains of a state.

Year	2001	2002	2003	2004	2005	2006
Production in 10000 tons	60	50	70	55	80	85

Choice of food for a group of people.

Favorite Food	**Number of people**
North Afghanistan	30
South Afghanistan	40
Chines	25
Others	25
Total	**120**

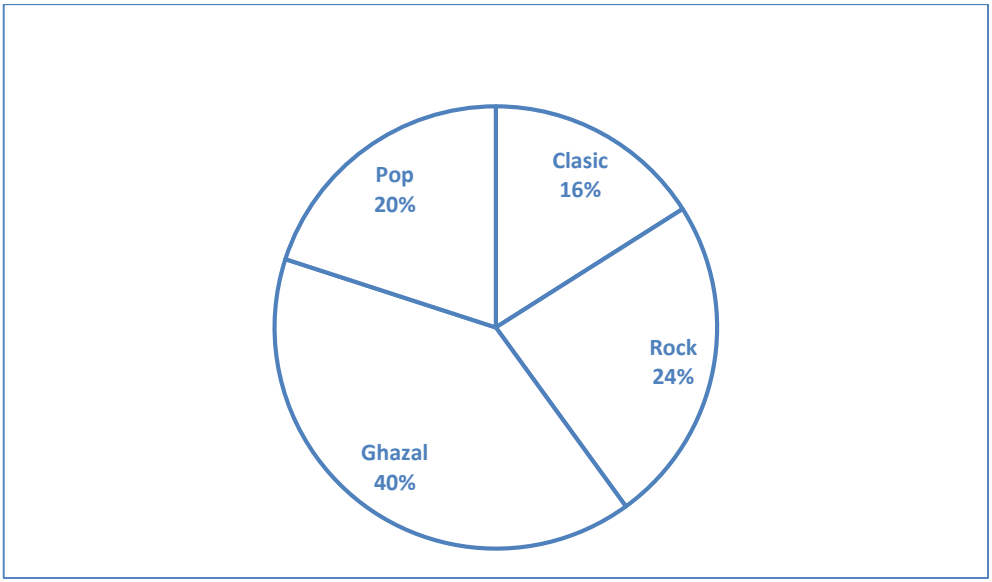
The daily income of a group of factory workers.

Daily income in Afs	**Number of workers in a factory**
75-100	45
100-125	35
125-150	55
150-175	30
175-200	50
200-225	125
225-250	140
Total	**480**

Chapter Exercise

A survey was made to find the type of music that as certain group of young people liked in a city. Adjoining pie chart shows the findings of this survey. From this pie chart answer the following:

a. If 20 people liked classical music, how many young people were surveyed?
b. Which type of music is liked by the maximum number of people?
c. If a cassette company were to make 1000 CD's, how many of each type would they make?



Draw a pie chart showing the following information. The table shows the colors preferred by a group of people.

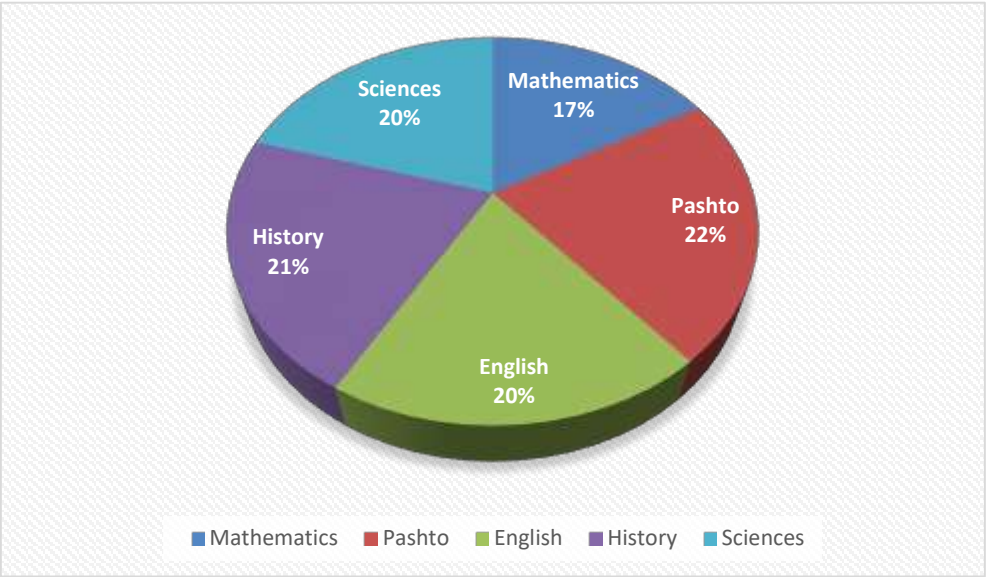
Colors	Preference
Red	20
Blue	10
Green	15
Black	30
White	20

The adjoining pie chart gives the marks scored in an examination by a student in Pashto, English, Mathematics, Arts and Science. If the total marks obtained by the student were 540, answer the following question.

(i) In which subject did the student score 105 marks?
(Hint: for 540 marks, the central angle = 360 degrees. So, for 105 marks, what is the central angle?)

(ii) How many more marks were obtained by the student in Mathematics than in History?

(iii) Examine whether the sum of the marks obtained in Arts and Mathematics is more than that in Science and Pashto.
(Hint: just study the central angles).



The number of students in a hostel, speaking different languages is given below. Display the data in a pie chart.

Language	English	Dari	Pashto	Arabic	French	Total
Number of students	40	12	30	8	4	94

Chapter Summary

1. We have seen that data is a collection of numbers gathered to give some information.
2. To get a particular information from the given data quickly, the data can be arranged in a tabular form using tally marks.
3. We learnt how a pictograph represents data in the form of pictures, objects or parts of objects. We have also seen how to interpret a pictograph and answer the related questions. We have drawn pictographs using symbols to represent a certain number of items or things. For example, ☺ = 100 students.
4. We discussed how to represent data by using a bar diagram or a bar graph. In a bar graph, bars of uniform width are drawn horizontally or vertically with equal spacing between them. The length of each bar gives the required information.
5. To do this we also discussed the process of choosing a scale for the graph. For example, 1 unit = 100 student. We have also practiced reading a given bar graph. We have seen how interpretations from the same can be made.

4

Measures of Central Tendency: Row Data

Introduction

In your previous lessons, you have dealt with various types of data. You have learnt to collect data, tabulate and put it in the form of bar graphs. The collection, recording and presentation of data helps us organize our experiences and draw inferences from them. In this Chapter, we will take one more step towards learning how to do this. You will come across some more kinds of data and graphs. You have seen several kinds of data through newspapers, magazines, television and other sources. You also know that all data give us some sort of information. Let us look at some common forms of data that you come across:

Table 1

Temperatures of Cities as on 20.06.2006		
CITY	**MAXIMUM**	**MINIMUM**
Ahmadabad	38^oC	29^oC
Amritsar	37^oC	26^oC
Bangalore	28^oC	21^oC
Chennai	36^oC	27^oC
Delhi	38^oC	28^oC
Jaipur	39^oC	29^oC
Jammu	41^oC	26^oC
Mumbai	32^oC	27^oC

Table 2

FOOTBALL WORLD CUP 2006	
Ukraine beat Saudi Arabia by	4-0
Spain beat Tunisia by	3-1
Switzerland beat Togo by	2-0

What do these collections of data tell you?

For example, you can say that the highest maximum temperature was in Jammu, in the table of temperatures of cities as on 20.06.2006 or you can say that the lowest minimum temperature was in Bangalore.

Can we organize and present these data in a different way, so that their analysis and interpretation becomes better? We shall address such questions in this Chapter.

Representative values

You might be aware of the term average and would have come across statements involving the term "average" in your day to day life:

- Helay spends on an average of about 5 hours daily for her studies.
- The average temperature at this time of the year is about 40 degree Celsius.
- The average age of pupils in Ebrahem's class is 12 years.
- The average attendance of students in a school during its final examination was 98 percent.

Many more of such statements could be there. Think about the statements given above. Do you think that the child in the first statement studies exactly for 5 hours daily? Or, is the temperature of the given place during that particular time always 40 degrees? Or, is the age of each pupil in that class 12 years? Obviously not. Then what do these statements tell you?

By average we understand that Helay, usually studies for 5 hours. On some days, she may study for less number of hours and on the other days she may study longer.

Similarly, the average temperature of 40 degrees Celsius, means that, very often the temperature at this time of the year is around 40 degree Celsius. Sometimes, it may be less than 40 degrees Celsius and at other times, it may be more than 40^oC.

Thus, we realize that average is a number that represents or shows the central tendency of a group of observations or data. Since average lies between the highest and the lowest value of the given data so, we say average is a measure of the central tendency of the group of data or an **average** is the sum of a list of numbers divided by the number of numbers in the list. Different forms of data need different forms of representative or central value to describe it. One of these representative values is the "Arithmetic mean". You will learn

about the other representative values in the remaining of the chapter. However, the word *average* may also refer to the median, mode, or other central or typical value. In statistics, these are all known as measures of central tendency.

Arithmetic mean

The most common representative value of a group of data is the arithmetic mean or the mean. To understand this in a better way, let us look at the following example:

Two vessels contain 20 liters and 60 liters of milk respectively. What is the amount that each vessel would have, if both share the milk equally? When we ask this question we are seeking the arithmetic mean.

In the above case, the average or the arithmetic mean would be

$$\frac{Total\ quantity\ of\ milk}{Number\ of\ vessels} = \frac{20+60}{2} liters = 40 liters.$$

Thus, each vessel would have 40 liters of milk.

The average or Arithmetic Mean (A. M.) or simply mean is defined as follows:

$$mean = \frac{Sum\ of\ all\ observations}{number\ of\ observations}$$

Consider these examples.

Example 1: Hameed studies for 4 hours, 5 hours and 3 hours respectively on three consecutive days. How many hours does he study daily on an average?

Solution: The average study time of Hameed would be

$$\frac{Total\ number\ of\ study\ hours}{Number\ of\ days\ for\ which\ he\ studied} = \frac{4+5+3}{3} hours = 4\ hours\ per\ day$$

Thus, we can say that Hameed studies for 4 hours daily on an average.

Example 2: A batsman scored the following number of runs in six innings:

36, 35, 50, 46, 60, 55

Calculate the mean runs scored by him in an inning.

Solution:

$Total\ runs = 36 + 35 + 50 + 46 + 60 + 55 = 282.$

To find the mean, we find the sum of all the observations and divide it by the number of observations.

Therefore, in this case, $mean = \frac{282}{6} = 47$. Thus, the mean runs scored in an inning are 47.

Where does the arithmetic mean lie?

TRY THESE

The arithmetic mean, often simply called the mean, of two numbers, such as 2 and 8, is obtained by finding a value A such that 2 + 8 = A + A. One may find that *A* = (2 + 8)/2 = 5.

Switching the order of 2 and 8 to read 8 and 2 does not change the resulting value obtained for A. The mean 5 is not less than the minimum 2 nor greater than the maximum 8. If we increase the number of terms in the list to 2, 8, and 11, the arithmetic mean is found by solving for the value of *A* in the equation 2 + 8 + 11 = *A* + *A* + *A*. One finds that *A* = (2 + 8 + 11)/3 = 7. How would you find the average of your study hours for the whole week?

THINK, DISCUSS AND WRITE

Consider the data in the above examples and think on the following:

- Is the mean bigger than each of the observations?
- Is it smaller than each observation?

Discuss with your friends. Frame one more examples of this type and answer the same questions.

You will find that the mean lies in between the greatest and the smallest observations.

In particular, the mean of two numbers will always lie between the two numbers. For example, the mean of 5 and 11 is $\frac{5+11}{2} = 8$, which lies between 5 and 11.

Can you use this idea to show that between any two fractional numbers, you can find as many fractional numbers as you like. For example, between $\frac{1}{2}$ and $\frac{1}{4}$ you have their average $\frac{\frac{1}{2}+\frac{1}{4}}{2} = \frac{3}{8}$ and then between $\frac{1}{2}$ and $\frac{3}{8}$, you have their average $\frac{7}{16}$ and so on.

TRY THESE

1. Find the mean of your sleeping hours during one week.
2. Find at least 5 numbers 5 numbers between $\frac{1}{2}$ and $\frac{1}{3}$.

The difference between the highest and the lowest observation gives us an idea of the spread of the observations. This can be found by subtracting the lowest observation from the highest observations. We call the result the range of the observation. Look at the following examples:

Example 3: The ages in years of 10 teachers of Karwan University are:

32, 41, 28, 54, 35, 26, 23, 33, 38, 40

a. What is the age of the oldest teacher and that of the youngest teacher?
b. What is the range of the ages of the teachers?
c. What is the mean age of these teachers?

1. Arranging the ages in ascending order, we get:

23, 26, 28, 32, 33, 35, 38, 40, 41, 54

We find that the age of the oldest teacher is 54 years and the age of the youngest teacher is 23.

2. Range of the ages of the teachers $= (54 - 23)\ years = 31\ years.$
3. Mean age of the teachers

$$\frac{23 + 26 + 28 + 32 + 33 + 35 + 38 + 40 + 41 + 54}{10}\ years = \frac{350}{10}\ years = 35\ years$$

EXERCISE 3.1

1. Find the range of heights of any ten students of your class.
2. Organize the following marks in a class assessment, in a tabular form.

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

a. Which number is the lowest?
b. Which number is the highest?
c. What is the range of the data?
d. Find the arithmetic mean.

3. Find the mean of the first five whole numbers.
4. A cricketer scores the following runs in eight innings:

58, 76, 40, 35, 46, 45, 0, 100.

Find the mean score.

5. Following table shows the points of each player scored in four games:
6.

Player	Game 1	Game 2	Game 3	Game 4
A	14	16	10	10
B	0	8	6	4
C	8	11	Did not play	13

Now answer the following questions:

a. Find the mean to determine A's average number of points scored per game.
b. To find the mean number of points per game for C, would you divide the total points by 3 or by 4? Why?
c. B played in all the four games. How would you find the mean?
d. Who is the best performer?

7. The marks (out of 100) obtained by a group of students in a science test are 85, 76, 90, 85, 39, 48, 56, 95, 81 and 75. Find the:

a. Highest and the lowest marks obtained by the students.
b. Range of the marks obtained.
c. Mean marks obtained by the group.

8. The enrolment in Karwan University during six consecutive years was as follows:

1555, 1670, 1750, 2013, 2540, 2820

Find the mean enrolment of the school for this period.

9. The rainfall (in mm) in a city on 7 days of a certain week was recorded as follows:
10.

Day	Mon	Tue	Wed	Thurs	Fri	Sat	Sun
Rainfall	0.0	12.2	2.1	0.0	20.5	5.5	1.0

a. Find the range of the rainfall in the above data.
b. Find the mean rainfall for the week.
c. On how many days was the rainfall less than the mean rainfall?

11. The highest of 10 girls were measured in cm and the results are as follows:
135, 150, 139, 128, 151, 132, 146, 149, 143, 141
    a. What is the height of the tallest girl?
    b. What is the height of the shortest girl?
    c. What is the range of the data?
    d. What is the mean height of the girls?
    e. How many girls have heights more than the mean height?

Mode

As we have said the mean is not the only measure of central tendency or the only form of representative value. For different requirements from a data, other measures of central tendencies are used.

To find out the weekly demand for different sizes of shirts, a shopkeeper kept records of sales of size $90\ cm$,$95\ cm$, $100\ cm$, $105\ cm$, $110\ cm$. Following is the record for a week:

Size (in inches)	$90\ cm$	$95\ cm$	$100\ cm$	$105\ cm$	$110\ cm$	Total
Number of shirts sold	8	22	32	37	6	105

If he found the mean number of shirts sold, do you think that he would be able to decide which shirt sizes to keep in stock?

$$Mean\ of\ total\ shirts\ sold = \frac{Total\ number\ of\ shirts\ sold}{Number\ of\ different\ sizes\ of\ shirts} = \frac{105}{5} = 21.$$

Should he obtain 21 shirts of each size? If he does so, will he be able to cater to the needs of the customers?

The shopkeeper, on looking at the record, decides to produce shirts of sizes $95\ cm$, $100\ cm$, $105\ cm$. He decided to postpone the procurement of the shirts of other sizes because of their small number of buyers.

The owner of a readymade dress shop says, The most popular size of dress I sell is the size $105\ cm$.

Observe that here also, the owner is concerned about the number of shirts of different sizes sold. She is however looking at the shirt size that is sold the most. This is another representative value for the data. The highest occurring event is the sale of size $90\ cm$. This representative value is called the mode of the data.

The mode of a set of observations is the observation that occurs most often. The most frequently occurring number in a list is called the mode. For example, the mode of the list (1, 2, 2, 3, 3, 3, 4) is 3. It may happen that there are two or more numbers which occur equally often and more often than any other number. In this case there is no agreed definition of mode. Some authors say they are all modes and some say there is no mode.

Find the mode of the given set of numbers: $1, 1, 2, 4, 3, 2, 1, 2, 2, 4$. Arrange the numbers with same values together, we get $1,1,1, 2, 2, 2, 2, 3, 4, 4$.

Mode of this data is 2 because it occurs more frequently than other observations. Putting the same observations together and counting them is not easy if the number of observations is large. In such cases we tabulate the data. Tabulation can begin by putting tally marks and finding the frequency, as you did in your previous lectures.

Look at the following example:

Example: Following are the margins of victory in the football matches of a league.

1,3,2,5,1,4,6,2,5,2,2,2,4,1,2,3,1,1,2,3,2,6,4,3,2,1,1,4,2,1,5,3,3,2,3,2,4,2,1,2.

Find the mode of this data.

Also find the mode of

a. 2,6,5,3,0,3,4,3,2,4,5,2,4,

b. 2,14,16,12,14,14,16,14,10,14,18,14.

Solution: Let us put the data in a tabular form:

Margins of Victory	Tally Bars	Number of Matches
1	(*IIIII*) *IIII*	9
2	(*IIIII*)(*IIIII*) *IIII*	14
3	(*IIIII*) *II*	7
4	(*IIIII*)	5
5	*III*	3
6	*II*	2
	Total	**40**

Looking at the table, we can quickly say that 2 is the 'mode' since 2 has occurred the highest number of times. Thus, most of the matches have been won with a victory margin of 2 goals.

THINK, DISCUSS AND WRITE

Can a set of numbers have more than one mode?

Example: Find the mode of the numbers 2,2,2,3,3,4,5,5,5,6,6,8.

Solution: Here, 2 and 5 both occur three times. Therefore, they both are modes of the data.

DO THIS

1. Record the age in years of all your classmates. Tabulate the data and find the mode.
2. Record the heights in centimeters of your classmates and find the mode.

TRY THESE

1. Find the mode of the following data:

12,14,12,16,15,13,14,18,19,12,14,15,16,15,16,16,15,17,13,16,16,15,15,
13,15,17,15,1415,13,15,14.

2. Heights (in cm) of 25 children are given below:

168,165,163,160,163,161,162,164,163,162,164,163,160,163,160,165,
163,162,163,164,163,160,165,163,162.

What is the mode of their heights? What do we understand by mode here?

Whereas mean gives us the average of all observations of the data, the mode gives that observation which occurs most frequently in the data.

Let us consider the following examples:

a. You have to decide upon the number of breads needed for 25 people called for a feast.
b. A shopkeeper selling shirts has decided to replenish her stock.
c. We need to find the height of the door needed in our house.
d. When going on a picnic, if only one fruit can be bought for everyone which is the fruit that we would get.

In which of these situations can we use the mode as a good estimate?

Consider the first statement. Suppose the number of breads needed by each person is 2,3,2,3,2,1,2,3,2,2,4,2,2,3,2,4,4,2,3,2,4,2,4,3,5.

The mode of the data is 2 breads. If we use mode as the representative value for this data, then we need 50 breads only, 2 for each of the 25 persons. However the total number would clearly be inadequate. Would mean be an appropriate representative value?

For the third statement the height of the door is related to the height of the persons using that door. Suppose there are 5 children and 4 adults using the door and the height of each of 5 children is around 135 cm. The mode for the heights is 135 cm. Should we get a door that is 144 cm high? Would all the adults be able to go through that door? It is clear that mode is not the appropriate representative value for this data. Would mean be an appropriate representative value here?

Why not? Which representative value of height should be used to decide the door height?

Similarly analyze the rest of the statements and find the representative value useful for that issue.

TRY THESE

Discuss with your friends and give

a. Two situations where mean would be an appropriate representative value to use, and
b. Two situations where mode would be an appropriate representative value to use.

Median

We have seen that in some situations, arithmetic mean is an appropriate measure of central tendency whereas in some other situations, mode is the appropriate measure of central tendency.

Let us now look at another example. Consider a group of 17 students with the following height (in cm):

106,110,123,125,117,120,112,115,110,120,115,102,115,115,109,115,101

The games teacher wants to divide the class into two groups so that each group has equal number of students, one group has students with height lesser than a particular height. How would she do that?

Let us see the various options she has:

1. She can find the mean. The mean is

$$\frac{106 + 110 + 123 + 125 + 117 + 120 + 112 + 115 + 110 + 120 + 115 + 102 + 115 + 115 + 109 + 115 + 101}{17}$$

$$= \frac{1930}{17} = 113.5$$

So, if the teacher divides the students into two groups on the basis of this mean height, such that one group has students of height less than the mean height and the other group has students with height more than the mean height, then the groups would be of unequal size. They would have 7 and 10 members respectively.

2. The second option for her is to find mode. The observation with highest frequency is $115\ cm$, which would be taken as mode.

There are 7 children below the mode and 10 children at the mode and above the mode. Therefore, we cannot divide the group into equal parts.

Let us therefore think of an alternative representative value or measure of central tendency. For doing this we again look at the given heights (in cm) of students arrange them in ascending order. We have the following observations:

$101,102,106,109,110,112,115,115,115,115,115,117,120,120,123,125$

The middle value in this data is 115 because this value divides the students into two equal groups of 8 students each. This value is called as Median. Median refers to the value which lies in the middle of the data (when arranged in an increasing or decreasing order) with half of the observations above it and the other half below it. The games teacher decides to keep the middle student as a referee in the game.

Here, we consider only those cases where number of observations is odd.

Thus, in a given data, arranged in ascending or descending order, the median gives us the middle observation.

In statistics, the median is the number separating the higher half of the data, from the lower half. The *median* of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one (e.g., the median of {3, 3, 5, 9, 11} is 5). If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values (the median of {3, 5, 7, 9} is (5 + 7) / 2 = 6), which corresponds to interpreting the median as the fully trimmed mid-range.

Chapter Exercise

Your friend found the median and the mode of a given data. Describe and correct your friend's error if any: $35, 32, 35, 42, 38, 32, 34$

$Median = 42, \quad Mode = 32.$

Note that in general, we may not get the same value for median and mode.

Thus we realize that mean, mode and median are the numbers that are the representative values of a group of observations or data. They lie between the minimum and maximum values of the data. They are also called the measures of the central tendency.

3. Find the median of the data: $24, 36, 46, 17, 18, 25, 35$.

Solution: We arrange the data in ascending order, we get $17, 18, 24, 25, 35, 36, 46$.
Median is the middle observation. Therefore 25 is the median.

1. The score in mathematics test (out of 25) of 15 students is as follows:
   $19, 25, 23, 20, 9, 20, 15, 10, 5, 16, 25, 20, 24, 12, 20$
   Find the mode and median of this data. Are they the same?
2. The runs scored in a cricket match by 11 players is as follows:
   $6, 15, 120, 50, 100, 80, 10, 15, 8, 10, 15$.
   Find the mean, mode and median of this data. Are the three same?
3. The weights (in kg) of 15 students of a class are:
   $38, 42, 35, 37, 45, 50, 32, 43, 43, 40, 36, 38, 43, 38, 47$.
   a. Find the mode and median of this data.
   b. Is there more than one mode?

4. Find the mode and median of the data: $13,16,12,14,19,12,14,13,14$.

5. Tell whether the statement is true or false:
   a. The mode is always one of the numbers in a data.
   b. The mean is one of the numbers in a data.
   c. The median is always one of the numbers in a data.
   d. The data $6, 4, 3, 8, 9, 12, 13, 9$ $has\ mean$ 9.

5

Measures of Central Tendency: Frequency Distribution

Earlier in chapter, we represented the data in various forms through frequency distribution table, bar graph, histograms and frequency polygons. Now, the question arises if we always need to study all the data to 'make sense' of it, or if we can make out some important features of it by considering only certain representatives of the data. This is possible, by using measures of central tendency or averages.

Consider a situation when two students Mary and Hadi received their test copies. The test had five questions, each carrying ten marks. Their scores were as follows:

Question Numbers	1	2	3	4	5
Mary's score	10	8	9	8	7
Hadi's score	4	7	10	10	10

Hadi said that since his middle most score was 10, which was higher than Mary's middle most score, that is 8, his performance should be rated better.

But Mary was not convinced. To convince Mary, Hadi tried out another strategy. He said he had scored 10 marks more often (3 times) as compared to Mary who scored 10 marks only once. So, his performance was better.

Now, to settle the dispute between Hadi and Mary, let us see the three measures they adopted to make their point.

Mean:

The average score that Mary found in the first case is the mean. The 'middle' score that Hari was using for his argument is the median. The most often scored mark that Hadi used in his second strategy is the mode.

Now, let us first look at the mean in detail.

The mean (or average) of a number of observations is the sum of the values of all the observations divided by the total number of observations. It is denoted by the symbol $\bar{x}$, read us 'x bar'.

Let us consider an example.

Example 10: 5 people were asked about the time in a week they spend in doing social work in their community. They said 10, 7, 13, 20 and 15 hours, respectively. Find the mean (or average) time in a week devoted by them for social work.

Solution: We have already studied in our earlier classes that the mean of a certain number of observations is equal to $\frac{Sum\ of\ all\ the\ observations}{Total\ number\ of\ observations}$. To simplify our working of finding the mean, let us use a variable x_1, second observation is x_2, and so on till x_5.

Also $x_1 = 10$ means that the value of the first observation, denoted by x_1, is 10. Similarly, $x_2 = 7, x_3 = 13, x_4 = 20\ and\ x_5 = 15$.

Therefore, the mean

$$\bar{x} = \frac{sum\ of\ all\ the\ observations}{total\ number\ of\ observations} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$= \frac{10 + 7 + 13 + 20 + 15}{5} = \frac{65}{5} = 13.$$

So, the mean time spent by these 5 people in doing social work is 13 hours in a week. Now, in case we are finding the mean time spent by 30 people in doing social work, writing $x_1 + x_2 + x_3 +, \cdots, +x_{30}$, we write $\sum_{i=1}^{30} x_i$, which is read as 'the sum of x_i as I varies from 1 to 30'.

So,

$$\bar{x} = \frac{\sum_{i=1}^{30} x_i}{30}$$

Similarly, for n observations:

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

Example 11: find the mean of the marks obtained by 30 students of a class, given in example 2.

Solution:

$$\bar{x} = \frac{x_1 + x_2 + x_3 +, \cdots, +x_{30}}{30},$$

$$\sum_{i=1}^{30} x_i = x_1 + x_2 + x_3 + , \cdots , + x_{30}$$

$$= 10 + 20 + 36 + 92 + 95 + 40 + 50 + 56 + 60 + 70 + 92 + 88 + 80 + 70 + 72 + 70 + 36 + 40 + 36 + 40 + 92 + 40 + 50 + 50 + 56 + 60 + 70 + 60 + 60 + 88 = 1779,$$

$$\bar{x} = \frac{1779}{30} = 59.3$$

Is the process not time consuming? Can we simplify it? Note that we have formed a frequency table for this data (see table 4.1)

The table shows that 1 student obtained 10 marks, 1 student obtained 20 marks, 3 students obtained 36 marks, 4 students obtained 40 marks, 3 students obtained 50 marks, 2 students obtained 56 marks, 4 students obtained 60 marks, 4 students obtained 70 marks, 1 student obtained 72 marks, 1 student obtained 80 marks, 2 students obtained 88 marks, 3 students obtained 92 marks and 1 student obtained 95 marks.

So, the total marks obtained

$$= (1 \times 10) + (1 \times 20) + (3 \times 36) + (4 \times 40) + (3 \times 50) + (2 \times 56) + (4 \times 60) + (4 \times 70) + (1 \times 72) + (1 \times 80) + (2 \times 88) + (3 \times 92) + (1 \times 95)$$

$$= f_1 x_1 + f_2 x_2 + , \cdots , + f_{13} x_{13},$$

Where f_i is the frequency of the i[th] entry in table 4.1?

In brief, we write this as $\sum_{i=1}^{13} f_i x_i$.

So, the total marks obtained

$$\sum_{i=1}^{13} f_1 x_i = 10 + 20 + 108 + 160 + 150 + 112 + 240 + 280 + 72 + 80 + 176 + 276 + 95 = 1779,$$

Now, the total number of observations:

$$\sum_{i=1}^{13} f_i = f_1 + f_2 + , \cdots , + f_{13} = 1 + 1 + 3 + 4 + 3 + 2 + 4 + 4 + 1 + 1 + 2 + 3 + 1 = 30,$$

So, the mean

$$\bar{x} = \frac{sum\ of\ all\ the\ observations}{total\ number\ of\ observations} = \frac{\sum_{i=1}^{13} f_i x_i}{\sum_{i=1}^{13} f_i} = \frac{1779}{30} = 59.3$$

This process can be displayed in the following table which is a modified form of table 4.1.

Marks (x_i)	Number of students (f_i)	$f_i x_i$
10	1	10
20	1	20
36	3	108
40	4	160
50	3	150
56	2	112

60	4	240
70	4	280
72	1	72
80	1	80
88	2	176
92	3	276
95	1	95
	$\sum_{i=1}^{13} f_i = 30$	$\sum_{i=1}^{13} f_i x_i = 1179$

Thus, in the case of an ungrouped frequency distribution, you can use the formula $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$ For calculating the mean.

Let us now move back to the situation of the argument between Hadi and Mary, and consider the second case where Hadi found his performance better by finding the middle-most score. As already stated, this measure of central tendency is called the median.

Median: is that value of the given number of observations, which divides it into exactly two parts. So, when the data is arranged in ascending (or descending) order the median of ungrouped data is calculated as follows:

a. When the number of observations (n) is odd, the median is the value of the $\left(\frac{n+1}{2}\right)^{th}$ observation. For example, if $n = 13$, the value of the $\left(\frac{13+1}{2}\right)^{th}$, i. e., the 7th observation will be the median.

b. When the number of observations (n) is even, the median is the mean of the $\left(\frac{n}{2}\right)^{th}$ and the $\left(\frac{n}{2}+1\right)^{th}$ observation. For example, if n=16, the mean of the value of the $\left(\frac{16}{2}\right)^{th}$ and the $\left(\frac{16}{2}+1\right)^{th}$ observations, i.e., the mean of the values of the 8th and 9th observations will be the median.

Let us illustrate this with the help of some examples.

Example 12: the height (in cm) of 9 students of a class are as follows:

155 160 145 149 150 147 152 144 148

Find the median of this data.

Solution: First of all, we arrange the data in ascending order, as follows:

144 145 147 148 149 150 152 155 160

Since the number of students is 9, an odd number, we find out the median by finding the height of the $\left(\frac{n+1}{2}\right)^{th} = \left(\frac{9+1}{2}\right)^{th} = the\ 5^{th}$ student, which is 149 cm. so, the median, i. e., the medial height is 149 cm.

Example 13: the points scored by a game team in a series of matches are as follows:

$$17,\ 2,\ 7,\ 27,\ 15,\ 5,\ 14,\ 8,\ 10,\ 24,\ 48,\ 10,\ 8,\ 7,\ 18,\ 28,$$

Find the median of the points scored by the team.

Solution: Arranging the points scored by the team in ascending order, we get:

$$2,\ 5,\ 7,\ 7,\ 8,\ 8,\ 10,\ 10,\ 14,\ 15,\ 17,\ 18,\ 24,\ 27,\ 28,\ 48,$$

There are 16 terms. So there are two middle terms, i. e. the $\frac{16}{2}$th and $\left(\frac{16}{2}+1\right) th, i.e.$, the 8^{th} and 9^{th} terms.

So, the median is the mean of the values of the 8^{th} and 9^{th} terms. i. e. the median $= \frac{10+14}{2} = 12$, so, the medial point scored by the sport team is 12.

Let us again go back to the unsorted dispute of Hadi and Mary.

Mode: The third measure used by Hadi to find the average was the mode.

The mode is that value of the observation which occurs most frequently, i. e., an observation with the maximum frequency is called the mode.

The readymade garment and shoes industries make great use of this measure of central tendency. Using the knowledge of mode, these industries decide which size of the product should be produced in large numbers.

Let us illustrate this with the help of an example.

Example 14: find the mode of the following marks (out of 10) obtained by 20 students:

$$4, 6, 5, 9, 3, 2, 7, 7, 6, 5, 4, 9, 10, 10, 3, 4, 7, 6, 9, 9,$$

Solution: We arrange this data in the following forms:

$$2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7, 7, 9, 9, 9, 9, 10, 10,$$

Here 9 occurs most frequently, i. e., four times. So, the mode is 9.

Example 15: consider a small unit of a factory where there are 5 employees. A supervisor and four laborers. The laborers draw a salary of 5000 Afs per month each while the supervisor gets 15000 Afs per month. Calculate the mean, median and mode of the salaries of this unit of factory.

Solution:

$$Mean = \frac{5000 + 5000 + 5000 + 5000 + 15000}{5} = \frac{35000}{5} = 7000,$$

So, the mean salary is 7000 Afs per month.

To obtain the median, we arrange the salaries in ascending order:

$$5000\ \ 5000\ \ 5000\ \ 5000\ \ 15000$$

Since the number of employees in the factory is 5, the median is given by the $\left(\frac{5+1}{2}\right)^{th} = \frac{6}{2} = 3^{rd}$ observation. Therefore, the median is Afs 5000 per month.

To find the mode of the salaries, i.e., the model salary, we see that 5000 occurs the maximum number of times in the data 5000, 5000, 5000, 5000, 15000. So, the modal salary is Afs 5000 per month.

Now compare the three measures of central tendency for the given data in the example above. You can see that the mean salary of Afs 7000 does not give even an approximate

estimate of any one of their wages, while the medial and modal salaries of Afs 5000 represents the data more effectively.

Extreme value in the data affect the mean. This in one of the weaknesses of the mean. So, if the data has a few points which are very far from most of the other points, (like 1, 7, 8, 9, 9) then the mean is not a good representative of this data. Since the median and mode are not affected by extreme values present in the data, they give a better estimate of the average in such a situation.

Again let us go back to the situation of Hadi and Mary, and compare the three measures of central tendency.

Measures of central tendency	Hadi	Mary
Mean	8.2	8.4
Median	10	8
Mode	10	8

This comparison helps us in stating that these measures of central tendency are not sufficient for concluding which student is better. We require some more information to conclude this, which you will study about in the higher classes.

Median of Data with Frequency Distribution

As you have studied in chapter four, the median is a measure of central tendency which gives the value of the middle-most observation in the data. Recall that for finding the median of ungrouped data, we first arrange the data values of the observations in ascending order. Then, if n is odd, the median is the $\left(\frac{n+1}{2}\right)^{th}$ observation. And, if n is even, then the median will be the average of the $\frac{n}{2}$ th and the $\left(\frac{n}{2}+1\right)^{th}$ observation.

Suppose, we have to find the median of the following data, which gives the marks, out of 50, obtained by 100 students in a test:

Marks Obtained	20	29	28	33	42	38	43	25
Number of students	6	28	24	15	2	4	1	20

First, we arrange the marks in ascending order and prepare a frequency table as follows:

Table 5.9.

Marks obtained	Number of students (Frequency)
20	6

25	20
28	24
29	28
33	15
38	4
42	2
43	1
Total	**100**

Here $n = 100$, which is even. The median will be the average of the $\frac{n}{2}\ th$ and the $\left(\frac{n}{2}+1\right)^{th}$ observation, i.e., the 50th and 51st observations. To find these observations, we proceed as follows:

Table 5.10.

Marks obtained	Number of students
20	6
Up to 25	$6 + 20 = 26$
Up to 28	$26 + 24 = 50$
Up to 29	$50 + 28 = 78$
Up to 33	$78 + 15 = 93$
Up to 38	$93 + 4 = 97$
Up to 42	$97 + 2 = 99$
Up to 43	$99 + 1 = 100$

Now we add another column depicting this information to the frequency table above and name it as cumulative frequency column.

Table 5.11.

Marks obtained	Number of students	Cumulative frequency
20	6	6
25	20	26
28	24	50
29	28	78
33	15	93
38	4	97
42	2	99
43	1	100

From the table above, we see that:

50th observation is 28 (why?)

51st observation is 29

So,

$$Median = \frac{28 + 29}{2} = 28.5$$

Remarks: The part of Table 5.11 consisting column 1 and column 3 is known as cumulative frequency table. The median marks 28.5 conveys the information that about 50% students obtained marks less than 28.5 and another 50% students obtained marks more than 28.5.

Mode of Grouped Data:

recall from chapter four, a mode is that value among the observations which occurs most often, that is, the value of the observation having the maximum frequency. Further, we discussed finding the mode of ungrouped data. Here, we shall discuss ways of obtaining a mode of grouped data. It is possible that more than one value may have the same maximum frequency. In such situations, the data is said to be multimodal. Though grouped data can also be multimodal, we shall restrict ourselves to problems having a single mode only.

Let us first recall how we found the mode for ungrouped data through the following example.

Example 4: the wickets taken by a bowler in 10 cricket matches are as follows:

2 6 4 5 0 2 1 3 2 3

Find the mode of the data.

Solution: Let us from the frequency distribution table of the given data as follows:

Number of wickets	0	1	2	3	4	5	6
Number of matches	1	1	3	2	1	1	1

Clearly, 2 is the number of wickets taken by the bowler in the maximum number (i.e.,) of matches. So, the mode of this data is 2.

Chapter Exercise

1. The following number of goals were scored by a team in a series of 10 matches:

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

Find the mean, median and mode of these scores.

2. In a mathematics test given to 15 students, the following marks (out of 100) are recorded:

$$41, 39, 48, 52, 46, 62, 54, 40, 96, 52, 98, 40, 42, 52, 60$$

Find the mean, median and mode of this data.

3. The following observations have been arranged in ascending order. If the median of the data is 63, find the value of x.

$$29, 32, 48, 50, x, x + 2, 72, 78, 84, 95$$

4. Find the modes of $14, 25, 14, 28, 18, 17, 18, 14, 23, 22, 14, 18$.
5. Find the mean salary of 60 workers of a factory from the following table:
   a. The mean is an appropriate measure of central tendency.
   b. The mean is not an appropriate measure of central tendency but the median

Salary (in Afs)	Number of workers
3000	16
4000	12
5000	10
6000	8
7000	6
8000	4
9000	3
10000	1
Total	**60**

is an appropriate measure of central tendency.

Chapter Summary

In this chapter, you have studied the following points:

1. Facts or figures, collected with a definite purpose, are called data.
2. Statistics is the area of study dealing with the presentation, analysis and interpretation of data.
3. How data can be presented graphically in the form of bar graph, histograms and frequency polygons.
4. The three measures of central tendency for ungrouped data are:

a. Mean: It is found by adding all the values of the observation and dividing it by the total number of observations. It is denoted by $\bar{x}$.

So, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$. For an ungrouped frequency distribution, it is $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$.

b. Median: It is the value of the middle most observation (s).

If n is an odd number, the median = value of the $\left(\frac{n+1}{2}\right)^{th}$ observation.

If n is an even number, median = mean of the value of the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observation.

c. Mode: the mode is the most frequently occurring observation.

6

Measures of Central Tendency: Grouped Data

In chapter four, you have studied the classification of given data into ungrouped as well as grouped frequency distributions. You have also learnt to represent the data pictorially in the form of various graphs such as bar graphs, histograms (including those of varying widths) and frequency polygons. In fact, you went a step further by studying certain numerical representatives of the ungrouped data, also called measures of central tendency, namely, mean, median, mode. In this chapter, we shall extend the study of these three measures, i.e., mean, median and mode from ungrouped data to that of grouped data. We shall also discuss the concept of cumulative frequency, the cumulative frequency distribution and how to draw cumulative frequency curves, called ogives.

Mean of Grouped Data:

The mean (or average) of observations, as we know, is the sum of the values of all the observations divided by the total number of observations. From chapter four, recall that if $x_1, x_2, x_3, \cdots, x_n$ are observations with respective frequencies $f_1, f_2, \cdots, f_n$, then this means observation x_1 occurs f_1 times, x_2 occurs f_2 times, and so on.

Now, the sum of the values of all the observations $= f_1x_1 + f_2x_2 +, \cdots, +f_nx_n$ and the number of observations $= f_1 + f_2 +, \cdots, +f_n$.

So, the mean $\bar{x}$ of the data is given by

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

Recall that we can write this in short form by using the Greek letter $\sum(capital\ sigma)$ which means summation. That is,

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

Which, more briefly, is written as $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$, if it is understood that i varies from 1 to n.

Let us apply this formula to find the mean in the following example.

Example 1: The marks obtained by 30 students of a class in a mathematics paper consisting of 100 marks are presented in table below. Find the mean of the marks obtained by the students.

Marks (x_i)	10	20	36	40	50	56	60	70	72	80	88	92	95
No of students (f_i)	1	1	3	4	3	2	4	4	1	1	2	3	1

Solution: Recall that to find the mean marks, we require the product of each x_i with the corresponding frequency f_i, so, let us put them in a column as shown in table 5.1.

Table 5.1.

Marks obtained (x_i)	**Number of students** (f_i)	$f_i x_i$
10	1	10
20	1	20
36	3	108
40	4	160
50	3	150
56	2	112
60	4	240
70	4	280
72	1	72
80	1	80
88	2	176
92	3	276
95	1	95
Total	$\sum f_i = 30$	$\sum f_i x_i = 1779$

Now,

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1779}{30} = 59.3$$

Therefore, the mean marks obtained is 59.3.

In most of our real life situations, data is usually so large that to make a meaningful study it needs to be condensed as grouped data. So, we need to convert given ungrouped data into grouped data and devise some method to find its mean.

Let us convert the ungrouped data of example 1 into grouped data by forming class intervals of width, say 15. Remember that, while allocating frequencies to each class interval, students falling in any upper class limit would be considered in the next class, e.g., 4 students who have obtained 40 marks would be considered in the class interval 40-55 and not in 25-40. With this convention in our mind, let us form a grouped frequency distribution table (see table 5.2).

Table 5.2.

Class interval	10-25	25-40	40-55	55-70	70-85	85-100
Number of students	2	3	7	6	6	6

Now, for each class interval, we require a point which would serve as the representative of the whole class. It is assumed that the frequency of each class interval is centered around its mid-point. So the mid-point (or class mark) of each class can be chosen to represent the observations falling in the class. Recall that we find the mid-point of a class (or its class mark) by finding the average of its upper and lower limits. That is,

$$Class\ Mark = \frac{Upper\ class\ limit + Lower\ class\ limit}{2}$$

With reference to table 5.2, for the class 10-25, the class mark is $\frac{10+25}{2}$, i.e., 17.5. Similarly, we can find the class marks of the remaining class intervals. We put them in table 5.3. These class marks serve as our x_i^s. Now, in general, for the i^{th} class interval, we have the frequency f_i corresponding to the class mark x_i. We can now proceed to compute the mean in the same manner as in Example 1.

Table 5.3.

Class interval	Number of students (f_i)	Class Mark (x_i)	$f_i x_i$
10-25	2	17.5	35.0
25-40	3	32.5	97.5
40-55	7	47.5	332.5
55-70	6	62.5	375.0
70-85	6	77.5	465.0
85-100	6	92.5	555.0
Total	$\sum \boldsymbol{f_i = 30}$		$\sum \boldsymbol{f_i x_i = 1860.0}$

The sum of the values in the last column gives us $\sum f_i x_i$. So, the mean $\bar{x}$ of the given data is given by

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1860.0}{30} = 62$$

This new method of finding the mean is known as the DIRECT METHOD.

We observe that Tables 5.1 and 5.3 are using the same data and employing the same formula for the calculation of the mean but the results obtained are different. Can you think

why this is so, and which one is more accurate? The difference in the two values is because of the mid-point assumption in Table 5.3, 59.3 being the exact mean, while 62 an approximate mean.

Sometimes when the numerical values of x_i and f_i are large, finding the product of x_i and f_i becomes tedious and time consuming. So, for each such situations, let us think of a method of reducing these calculations.

We can do nothing with the f_i's, but we can change each x_i to a smaller number so that our calculations become easy. How do we do this? What about subtracting a fixed number from each of these $x_i's$? Let us try this method.

The first step is to choose one among the $x_i's$ as the assumed mean, and denote it by 'a'. Also, to further reduce our calculation work, we may take 'a' to be that x_i which lies in the center of $x_1, x_2, \cdots, x_n$. So, we can choose $a = 47.5$ *or* $a = 62.5$. let us choose $a = 47.5$.

The next step is to find the difference d_i between a and each of the $x_i's$, that is, the deviation of 'a' from each of the $x_i's$.

i.e.,

$$d_i = x_i - a = x_i - 47.5$$

The third step is to find the product of d_i with the corresponding f_i, and take the sum of all the $f_i d_i$ s. the calculations are shown in Table 5.4.

Table 5.4.

Class interval	**Number of students (f_i)**	**Class mark (x_i)**	$\boldsymbol{d_i = x_i - 47.5}$	$\boldsymbol{f_i d_i}$
10-25	2	17.5	-30	-60
25-40	3	32.5	-15	-45
40-55	7	47.5	0	0
55-70	6	62.5	15	90
70-85	6	77.5	30	180
85-100	6	92.5	45	270
Total	$\boldsymbol{\sum f_i = 30}$			$\boldsymbol{\sum f_i d_i = 435}$

So, from Table 5.4, the mean of the deviations,

$$\bar{d} = \frac{\sum f_i d_i}{\sum f_i}$$

Now, let us find the relation between $\bar{d}$ and $\bar{x}$.

Since in obtaining d_i, we subtracted 'a' from each $x_i's$, so, in order to get the mean $\bar{x}$, we need to add 'a' to $\bar{d}$. This can be explained mathematically as:

Mean of deviations,

$$\bar{d} = \frac{\sum f_i d_i}{\sum f_i}$$

So,

$$\bar{d} = \frac{\sum f_i (x_i - a)}{\sum f_i} = \frac{\sum f_i x_i}{\sum f_i} - \frac{\sum f_i a}{\sum f_i} = \bar{x} - a\frac{\sum f_i}{\sum f_i} = \bar{x} - a$$

So,

$$\bar{x} = a + \bar{d}$$

i.e.,

$$\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i}$$

Substituting the value of a, $\sum f_i d_i$ and $\sum f_i$ from Table 5.4, we get

$$\bar{x} = 47.5 + \frac{435}{30} = 47.5 + 14.5 = 62.$$

Therefore, the mean of the marks obtained by the students is 62.

The method discussed above is called the **Assumed Mean Method.**

Activity 1: From the Table 5.3 find the mean by taking each of x_i (i. e. 17.5, 32.5, and so on) as 'a'. what do you observe? You will find that the mean determined in each case is the same, i. e., 62. (why?)

So, we can say that the value of the mean obtained does not depend on the choice of 'a'.

Observe that in Table 5.4, the value in column 4 are all multiple of 15. So, if we divide the values in the entire column 4 by 15, we would get smaller numbers to multiply with f_i. (here, 15 is the class size of each class interval.)

So, let $u_i = \frac{x_i - a}{h}$, where a is the assumed mean and h is the class size.

Now, we calculate u_i in this way and continue as before (i. e., find $f_i u_i$ *and* the $\sum f_i u_i$).

Taking $h = 15$, *let us form Table* 5.5.

Table 5.5

Class interval	f_i	x_i	$d_i = x_i - a$	$u_i = \frac{x_i - a}{h}$	$f_i u_i$
10-25	2	17.5	-30	-2	-4
25-40	3	32.5	-15	-1	-3
40-55	7	47.5	0	0	0
55-70	6	62.5	15	1	6
70-85	6	77.5	30	2	12
85-100	6	92.5	45	3	18
Total	$\sum f_i = 30$				$\sum f_i u_i = 29$

Let

$$\bar{u} = \frac{\sum f_i u_i}{\sum f_i}$$

Here, again let us find the relation between $\bar{u}$ and $\bar{x}$.

We have,

$$u_i = \frac{x_i - a}{h}$$

Therefore,

$$\bar{u} = \frac{\sum f_i \frac{(x_i - a)}{h}}{\sum f_i} = \frac{1}{h}\left[\frac{\sum f_i x_i - a\sum f_i}{\sum f_i}\right] = \frac{1}{h}\left[\frac{\sum f_i x_i}{\sum f_i} - a\frac{\sum f_i}{\sum f_i}\right] = \frac{1}{h}[\bar{x} - a],$$

So,

$$h\bar{u} = \bar{x} - a$$

i.e.,

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

So,

$$\bar{x} = a + h\left(\frac{\sum f_i u_i}{\sum f_i}\right)$$

Now, substituting the value of a, h, $\sum f_i u_i$ and $\sum f_i$ from Table 5.5, we get

$$\bar{x} = 47.5 + 15 \times \left(\frac{29}{30}\right) = 47.5 + 14.5 = 62.$$

So, the mean marks obtained by a student is 62.

The method discussed above is called the step deviation method.

We note that:

- The step deviation method will be convenient to apply if all the d_is have a common factor.
- The mean obtained by all the three methods is the same.
- The assumed mean method and step deviation method are just simplified forms of direct method.
- The formula $\bar{x} = a + h\bar{u}$ still hold if a and h are not as given above, but are any non zero numbers such that $u_i = \frac{x_i - a}{h}$.

Let us apply these methods in another example.

Example 2: the table below gives the percentage distribution of female teachers in the primary schools of rural areas of various states and union territories (U.T.) of India. Find the mean percentage of female teachers by all the three methods discussed in this section.

Female Teachers (%)	15-25	25-35	35-45	45-55	55-65	65-75	75-85
(U.T.)	6	11	7	4	4	2	1

Source: Seventh All India School Education Survey conducted by NCERT.

Solution: Let us find the class marks, x_i, of each class, and put them in a column (see Table 5.6):

Table 5.6.

Female Teachers (%)	**(U.T.)**	x_i
15-25	6	20
25-35	11	30
35-45	7	40

45-55	4	50
55-65	4	60
65-75	2	70
75-85	1	80

Here we take $a = 50, h = 10$ then $d_i = x_i - 50$ and $u_i = \frac{x_i - 50}{10}$.

We now find d_i and u_i and put them in Table 5.7.

Female Teachers (%)	**(U.T.)** (f_i)	x_i	$d_i = x_i - 50$	$u_i = \frac{x_i - 50}{10}$	$f_i x_i$	$f_i d_i$	$f_i u_i$
15-25	6	20	-30	-3	120	-180	-18
25-35	11	30	-20	-2	330	-220	-22
35-45	7	40	-10	-1	280	-70	-7
45-55	4	50	0	0	200	0	0
55-65	4	60	10	1	240	40	4
65-75	2	70	20	2	140	40	4
75-85	1	80	30	3	80	30	3
Total	**35**				**1390**	**-360**	**-36**

From the table above, we obtain

$$\sum f_i = 35, \quad \sum f_i x_i = 1390, \quad \sum f_i d_i = -360, \quad \sum f_i u_i = -36,$$

Using the direct method, we get:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1390}{35} = 39.71,$$

Using the assumed mean method, we get:

$$\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i} = 50 + \frac{(-360)}{35} = 39.71,$$

Using the step – deviation method we get:

$$\bar{x} = a + \left(\frac{\sum f_i u_i}{\sum f_i}\right) \times h = 50 + \left(\frac{-36}{35}\right) \times 10 = 39.71,$$

Therefore, the mean percentage of female teachers in the primary schools of rural areas is 39.71.

Remark: the result obtained by all the three methods is the same. So the choice of method to be used depends on the numerical values of x_i and f_i. If x_i and f_i are numerically large numbers, then we can go for the assumed mean method or step-deviation method. If the class sizes are unequal, and x_i are large numerically, we can still apply the step- deviation method by taking h to be a suitable divisor of all the d_is.

Example 3: The distribution below shows the number of wickets taken by bowlers in one-day cricket matches. Find the mean number of wickets by choosing a suitable method. What does the mean signify?

No of Wickets	20-60	60-100	100-150	150-250	250-350	350-450
No of Bowlers	7	5	16	12	2	3

Solution: Here, the class size varies, and the x_is are large. Let us still apply the step-deviation method with $a = 200\ and\ h = 20$. Then, we obtain the data as in Table 5.8.

Table 5.8.

No of Wickets	**No of Bowlers** (f_i)	x_i	$d_i = x_i - 200$	$u_i = \frac{d_i}{20}$	$u_i f_i$
20-60	7	40	-160	-8	-56
60-100	5	80	-120	-6	-30
100-150	16	125	-75	-3.75	-60
150-250	12	200	0	0	0
250-350	2	300	100	5	10
350-450	3	400	200	10	30
Total	**45**				**-106**

So, $\bar{u} = \frac{-106}{45}$. Therefore, $\bar{x} = 200 + 20\left(\frac{-106}{45}\right) = 200 - 47.11 = 152.89$.

This tells us that, on an average, the number of wickets taken by these 45 bowlers in one-day cricket match is 152.89.

Now, let us see how well you can apply the concepts discussed in this section!

Activity 2: divide the students of your class into three groups and ask each group to do one of the following activities.

1. Collect the marks obtained by all the students of your class in mathematics in the latest examination conducted by your school. Form a grouped frequency distribution of the data obtained.
2. Collect the daily maximum temperatures recorded for a period of 30 days in your city. Present this data as a grouped frequency table.
3. Measure the heights of all the students of your of your class (in cm) and form a grouped frequency distribution table of this data.

After all the groups have collected the data and formed grouped frequency distribution table, the groups should find the mean in each case by the method which they find appropriate.

EXERCISE 5.1.

1. A survey was conducted by a group of students as a part of their environment awareness program, in which they collected the following data regarding the number of plants in 20 houses in a locality. Find the mean number of plants per house.

No of plants	0-2	2-4	4-6	6-8	8-10	10-12	12-14
No of houses	1	2	1	5	6	2	3

Which method did you use for finding the mean, and why?

2. Consider the following distribution of daily wages of 50 workers of a factory.

Daily wages (in Afs)	100-120	120-140	140-160	160-180	180-200
Number of workers	12	14	8	6	10

Find the mean daily wages of the workers of the factory by using an appropriate method.

3. The following distribution shows the daily pocket allowance of children of a locality. The mean pocket allowance is Afs 18. Find the missing frequency f.

Daily pocket allowance (in Afs)	11-13	13-15	15-17	17-19	19-21	21-23	23-25
Number of children	7	6	9	13	f	5	4

4. Thirty women were examined in a hospital by a doctor and the number of heart beats per minute were recorded and summarized as follows. Find the mean heart beat s per minute for these women, choosing a suitable method.

Number of heart beats per minute	65-68	68-71	71-74	74-77	77-80	80-83	83-86
Number of women	2	4	3	8	7	4	2

5. In a retail market, fruit vendors were selling mangoes kept in packing boxes. These boxes contained varying number of mangoes. The following was the distribution of mangoes accourding to the number of boxes.

Number of Mangoes	50-52	53-55	56-58	59-61	62-64
Number of Boxes	15	110	135	115	25

Find the mean number of mangoes kept in a packing box. Which methods of finding the mean did you choose?

6. The table below shows the daily expenditure on food of 25 households in a locality.

Daily expenditure (in Afs)	100-150	150-200	200-250	250-300	300-350
Number of households	4	5	12	2	2

Find the mean daily expenditure on food by a suitable method.

7. To find out the concentration of SO_2 in the air (in parts per million, i. e., ppm), the data was collected for 30 localities in a certain city and is presented below:

Concentration of SO_2 (in ppm)	**Frequency**
0.00-0.04	4
0.04-0.08	9
0.08-0.12	9
0.12-0.16	2
0.16-0.20	4
0.20-0.24	2

Find the mean concentration of SO_2 in the air.

8. A class teacher has the following absentee record of 40 students of a class for the whole term. Find the mean number of days a student was absent.

No of Days	0-6	6-10	10-14	14-20	20-28	28-38	38-40
No of students	11	10	7	4	4	3	1

9. The following table gives the literacy rate (in percentage) of 35 cities. Find the mean literacy rate.

Literacy rate (%)	45-55	55-65	65-75	75-85	85-95
Number of cities	3	10	11	8	3

Mode of Grouped Data:

recall from chapter four, a mode is that value among the observations which occurs most often, that is, the value of the observation having the maximum frequency. Further, we discussed finding the mode of ungrouped data. Here, we shall discuss ways of obtaining a mode of grouped data. It is possible that more than one value may have the same maximum frequency. In such situations, the data is said to be multimodal. Though grouped data can also be multimodal, we shall restrict ourselves to problems having a single mode only.

Let us first recall how we found the mode for ungrouped data through the following example.

Example 4: the wickets taken by a bowler in 10 cricket matches are as follows:

2 6 4 5 0 2 1 3 2 3

Find the mode of the data.

Solution: Let us from the frequency distribution table of the given data as follows:

Number of wickets	0	1	2	3	4	5	6
Number of matches	1	1	3	2	1	1	1

Clearly, 2 is the number of wickets taken by the bowler in the maximum number (i.e.,) of matches. So, the mode of this data is 2.

In a grouped frequency distribution, it is not possible to determine the mode by looking at the frequencies. Here, we can only locate a class with the maximum frequency, called the modal class. The mode is a value inside the modal class, and is given by the formula:

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

Where

$l = lower\ limit\ of\ the\ modal\ class,$

$h = size\ of\ the\ class\ interval\ (assuming\ all\ class\ sizes\ to\ be\ equal),$

$f_1 = frequency\ of\ the\ modal\ class,$

$f_0 = frequency\ of\ the\ class\ preceding\ the\ modal\ class,$

$f_2 = frequency\ of\ the\ class\ succeeding\ the\ modal\ class.$

Let us consider the following examples to illustrate the use of this formula.

Example 5: A survey conducted on 20 households in a locality by a group of students resulted in the following frequency table for the number of family members in a household:

Family size	1-3	3-5	5-7	7-9	9-11
Number of Families	7	8	2	2	1

Find the mode of this data.

Solution: Here the maximum class frequency is 8, and the class corresponding to this frequency is 3-5. So, the modal class is 3-5.

Now

$modal\ class = 3-5, lower\ limit\ (l)\ of\ modal\ class = 3,$
$class\ size\ (h) = 2, frequency\ (f_1)\ of\ the\ modal\ class = 8,$
$frequency\ (f_0)\ of\ class\ preceding\ the\ modal\ class = 7,$
$frequency\ (f_2)\ of\ class\ succeeding\ the\ modal\ class = 2.$

Now, let us substitute these values in the formula:

$$Mode = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h = 3 + \left(\frac{8-7}{2\times 8 - 7 - 2}\right) \times 2 = 3 + \frac{2}{7} = 3.286$$

Therefore, the mode of the data above is 3.286.

Example 6: The marks distribution of 30 students in a mathematics examination are given in Table 5.3 of Example 1. Find the mode of this data. Also compare and interpret the mode and the mean.

Solution: Refer to Table 5.3 of Example 1. Since the maximum number of students (i.e., 7) have got marks in the interval 40-55, the modal class is 40-55. Therefore, the lower limit (l) of the modal class=40,

The class size $(h) = 15$,

The frequency (f_1) of modal class=7,

The frequency (f_0) of the class preceding the modal class=3,

The frequency (f_2) of the class succeeding the modal class=6.

Now, using the formula:

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

We get

$$Mode = 40 + \left(\frac{7-3}{14-6-3}\right) \times 15 = 52$$

So, the mode marks is 52.

Now, from Example 1, you know that the mean marks is 62.

So, the maximum number of students obtained 52 marks, while on an average a student obtained 62 marks.

Remarks:

1. In Example 6, the mode is less than the mean. But for some other problems in may be equal or more than the mean also.
2. It depends upon the demand of the situation whether we are interested in finding the average marks obtained by the students or the average of the marks obtained

by most of the students. In the first situation, the mean is required and in the second situation, the mode is required.

Activity 3: Continuing with the same groups as formed in Activity 2 and the situations assigned to the groups. Ask each group to find the mode of the data. They should also compare this with the mean, and interpret the meaning of both.

Remarks: The mode can also be calculated for grouped data with unequal class sizes. However, we shall not be discussed it.

EXERCISE 5.2.

1. The following table shows the ages of the patients admitted in a hospital during a year:

Ages (in years)	5-15	15-25	25-35	35-45	45-55	55-65
Number of patients	6	11	21	23	14	5

Find the mode and the mean of the data given above. Compare and interpret the two measures of central tendency.

2. The following data gives the information on the observed lifetimes (in hours) of 225 electrical components:

Lifetimes (in hours)	0-20	20-40	40-60	60-80	80-100	100-120
Frequency	10	35	52	61	38	29

Determine the modal lifetimes of the components.

3. The following data gives the distribution of total monthly household expenditure of 200 families of a village. Find the modal monthly expenditure of the families. Also, find the mean monthly expenditure:

Expenditure (in Afs)	**Number of Families**
1000-1500	24
1500-2000	40
2000-2500	33
2500-3000	28
3000-3500	30
3500-4000	22
4000-4500	16
4500-5000	7

4. The following distribution gives the state-wise teacher-student ratio in higher secondary schools on India. Find the mode and mean of this data. Interpret the two measures.

Number of students per teacher	Number of states / U.T.
15-20	3
20-25	8
25-30	9
30-35	10
35-40	3
40-45	0
45-50	0
50-55	2

5. The given distribution shows the number of runs scored by some top batsman of the world in one-day international cricket matches.

Runs Scored	Number of Batmen
3000-4000	4
4000-5000	18
5000-6000	9
6000-7000	7
7000-8000	6
8000-9000	3
9000-10000	1
10000-11000	1

Find the mode of the data.

6. A student noted the number of cars passing through a spot on a road for 100 periods each of 3 minutes and summarized it in the table given below. Find the mode of the data:

Number of cars	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	7	14	13	12	20	11	15	8

Median of Grouped Data:

Now, let us see how to obtain the median of grouped data, through the following situation. Consider a grouped frequency distribution of marks obtained, out of 100, by 53 students, in a certain examination, as follows:

Table 5.12.

Marks	Number of students (Frequency)
0-10	5
10-20	3
20-30	4
30-40	3

40-50	3
50-60	4
60-70	7
70-80	9
80-90	7
90-100	8

From the table above, try to answer the following questions:

How many students have scored marks less than 10? The answer is clearly 5.

How many students have scored less than 20 marks? Observe that the number of students who have scored less than 20 include the number of students who have scored marks from 0-10 as well as the number of students who have scored marks from 10-20. So, the total number of students with marks less than 20 is 5+3, i.e., 8. We say that the cumulative frequency of the class 10-20 is 8.

Similarly, we can compute the cumulative frequencies of the other classes, i.e., the number of students with marks less than 30, less than 40, …, less than 100. We give them in Table 5.13 given below:

Table 5.13.

Marks obtained	**Number of students (Cumulative frequency)**
$less\ than\ 10$	5
$less\ than\ 20$	$5+3=8$
$less\ than\ 30$	$8+4=12$
$less\ than\ 40$	$12+3=15$
$less\ than\ 50$	$15+3=18$
$less\ than\ 60$	$18+4=22$
$less\ than\ 70$	$22+7=29$
$less\ than\ 80$	$29+9=38$
$less\ than\ 90$	$38+7=45$
$less\ than\ 100$	$45+8=53$

The distribution given above is called the cumulative frequency distribution of the less than type. Here 10, 20, 30, …, 100, are the upper limits of the respective class intervals.

We can similarly make the table for the number of students with scores, more than or equal to 0, more than or equal to 10, more than or equal to 20, and so on. From Table 5.12, we observe that all 53 students have scored marks more than or equal to 0. Since there are 5 students scoring marks in the interval 0-10, this mean that there are 53-5=48 students getting more than or equal to 10 marks. Continuing in the same manner, we get the number of students scoring 20 or above as 48-3=45. 30 or above as 45-4=41, and so on, as shown in Table 5.14.

Table 5.14.

Marks obtained	**Number of students (Cumulative frequency)**
$more\ than\ or\ equal\ to\ 0$	53

more than or equal to 10	$53 - 5 = 48$
more than or equal to 20	$48 - 3 = 45$
more than or equal to 30	$45 - 4 = 41$
more than or equal to 40	$41 - 3 = 38$
more than or equal to 50	$38 - 3 = 35$
more than or equal to 60	$35 - 4 = 31$
more than or equal to 70	$31 - 7 = 24$
more than or equal to 80	$24 - 9 = 15$
more than or equal to 90	$15 - 7 = 8$

The table above is called a cumulative frequency distribution of the more than type. Here 0, 10, 20, …, 90 give the lower limits of the respective class intervals.

Now, to find the median of grouped data, we can make use of any of these cumulative frequency distributions.

Let us combine Tables 5.12 and 5.13 to get Table 5.15 given bellow:

Table 5.15.

Marks	Number of students (f)	Cumulative Frequency (cf)
0-10	5	5
10-20	3	8
20-30	4	12
30-40	3	15
40-50	3	18
50-60	4	22
60-70	7	29
70-80	9	38
80-90	7	45
90-100	8	53

Now in a grouped data, we may not be able to find the middle observation by looking at the cumulative frequencies as the middle observation will be some value in a class interval. It is, therefore, necessary to find the value inside a class that divides the whole distribution into two halves. But which class should this be?

To find this class, we find the cumulative frequencies of all the classes and $\frac{n}{2}$. We now locate the class whose cumulative frequencies is greater than (and nearest to) $\frac{n}{2}$. This is called the median class. In the distribution above, $n = 53$. So, $\frac{n}{2} = 26.5$. now $60 - 70$ is the class whose cumulative frequency 29 is greater than (and nearest to) $\frac{n}{2}$, i.e., 26.5.

Therefore, $60 - 70$ is the median class.

After finding the median class, we use the following formula for calculating the median.

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

Where
$l = lower\ limit\ of\ median\ class,$
$n = number\ of\ observations,$
$cf = cumulative\ frequency\ of\ class\ preceding\ the\ median\ class,$
$f = frequency\ of\ median\ class,$
$h = class\ size\ (assuming\ class\ size\ to\ be\ equal).$
Substituting the value $\frac{n}{2} = 26.5, l = 60, cf = 22, f = 7, h = 10$ in the formula above, we get

$$Median = 60 + \left(\frac{26.5 - 22}{7}\right) \times 10 = 60 + \frac{45}{7} = 66.4$$

So, about half the students have scored marks less than 66.4, and the other half have scored marks more than 66.4.
Example 7: A survey regarding the heights (in cm) of 51 students of a class was conducted and the following data was obtained:

Height (in cm)	Number of students
$less\ than\ 140$	4
$less\ than\ 145$	11
$less\ than\ 150$	29
$less\ than\ 155$	40
$less\ than\ 160$	46
$less\ than\ 165$	51

Find the median height.
Solution: To calculate the median height, we need to find the class interval and their corresponding frequencies.
The given distribution being of the less than type, 140, 145, 150, …, 165 give the upper limits of the corresponding class intervals. So, the classes should be below 140, 140-145, 145-150, …, 160-165. Observe that from the given distribution, we find that there are 4 students with height less than 140, i.e., the frequency of class interval below 140 is 4. Now, there are 11 students with heights less than 145 and 4 students with height less than 140. Therefore, the number of students with height in the interval 140-145 is 11-4=7. Similarly, the frequency of 145-150 is 29-11=18, for 150-155, it is 40-29=11, and so on. So, our frequency distribution table with the given cumulative frequencies becomes:

Table 5.16.

Class Intervals	Frequency	Cumulative Frequency
$below\ 140$	4	4
$140 - 145$	7	11

$145-150$	18	29
$150-155$	11	40
$155-160$	6	46
$160-165$	5	51

Now $n = 51$. So, $\frac{n}{2} = 25.5$. this observation lies in the class 145-150. Then,
$l\ (the\ lower\ limit) = 145$,
$cf\ (the\ cumulative\ frequency\ of\ the\ class\ preceding\ 145-150) = 11$,
$f(the\ frequency\ of\ the\ median\ class\ 145-150) = 18$,
$h\ (the\ class\ size) = 5$.

Using the formula, $Median = l + \left(\frac{\frac{n}{2}-cf}{f}\right) \times h$, we have

$$Median = 145 + \left(\frac{25.5-11}{18}\right) \times 5 = 145 + \frac{72.5}{18} = 149.03.$$

So, the median height of the students is 149.03 cm.

This means that the height of about 50% of the students is less than this height, and 50% are taller than this height.

Example 8: The median of the following data is 525. Find the value of x and y, if the total frequency is 100.

Class interval	Frequency
0-100	2
100-200	5
200-300	x
300-400	12
400-500	17
500-600	20
600-700	y
700-800	9
800-900	7
900-1000	4

Solution:

Class intervals	Frequency	Cumulative Frequency
0-100	2	2
100-200	5	7
200-300	x	$7 + x$
300-400	12	$19 + x$
400-500	17	$36 + x$
500-600	20	$56 + x$

600-700	y	$56+x+y$
700-800	9	$65+x+y$
800-900	7	$72+x+y$
900-1000	4	$76+x+y$

It is given that n=100.

So, $76+x+y=100, i.e., x+y=24$ (i)

The median is 525, which lies in the class 500-600.

So, $l=500, f=20, cf=36+x, h=100,$

Using the formula:

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

$$525=500+\left(\frac{50-36-x}{20}\right)\times 100$$

$$525-500=(14-x)\times 5$$

$$25=70-5x$$

$$5x=70-25=45$$

$$x=9$$

Therefore, from (i), we get $9+y=24$

$$y=15$$

Now, that you have studied about all the three measures of central tendencies, let us discuss which measure would be best suited for a particular requirement.

The mean is the most frequently used measure of central tendency because it takes into account all the observations, and lies between the extremes, i.e., the largest and the smallest observations of the entire data. It also enables us to compare two or more distributions. For example, by comparing the average (mean) results of students of different schools of a particular examination, we can conclude which school has a better performance.

However, extreme values in the data affect the mean. For example, the mean of classes having frequencies more or less the same is a good representative of the data. But, if one class has frequency, say 2, and the five others have frequency 20, 25, 20, 21, 18, then the mean will certainly not reflect the way the data behaves. So, in such cases, the mean is not a good representative of the data.

In problems where individual observations are not important, and we wish to find out a 'typical' observation, the median is more appropriate, e.g., finding the typical productivity rate of workers, average wage in a country, etc. these are situations where extreme values may be there. So, rather than the mean, we take the median as a better measure of central tendency.

In situations which required establishing the most frequent value or most popular item, the mode is the best choice, e.g., to find the most popular T.V. program being watched, the consumer item in greatest demand, the color of the vehicle used by most of the people, etc.
Remark:

1. There is empirical relationship between the three measures of central tendency:
$$3\,Median = Mode + 2\,Mean$$
2. The median of grouped data with unequal class sizes can also be calculated. However, we shall not discuss it here.

EXERCISE 5.3.

1. The following frequency distribution gives the monthly consumption of electricity of 68 consumers of a locality. Find the median, mean and mode of the data and compare them.

Monthly consumption (in units)	Number of consumers
65-85	4
85-105	5
105-125	13
125-145	20
145-165	14
165-185	8
185-205	4

2. If the median of the distribution given below is 28.5, find the value of x and y.

Class interval	Frequency
0-10	5
10-20	x
20-30	20
30-40	15
40-50	y
50-60	5
Total	60

3. A life insurance agent found the following data for distribution of ages of 100 policy holders. Calculate the median age, if policies are given only to persons having age 18 years onwards but less than 60 years.

Age (in years)	Number of policy holders
$Below\ 20$	2
$Below\ 25$	6
$Below\ 30$	24
$Below\ 35$	45
$Below\ 40$	78

$Below\ 45$	89
$Below\ 50$	92
$Below\ 55$	98
$Below\ 60$	100

4. The length of 40 leaves of a plant are measured correct to the nearest millimeters, and the data obtained is represented in the following table:

Length (in mm)	**Number of leaves**
118-126	3
127-135	5
136-144	9
145-153	12
154-162	5
163-171	4
172-180	2

Find the median length of the leaves.
(Hint: The data needs to be converted to continues classes for finding the median, since the formula assumes continuous classes. The classes then change to 117.5-126.5, 126.5-135.5, …, 171.5-180.5)

5. The following table gives the distribution of the life time of 400 neon lamps:

Life time (in hours)	**Number of lamps**
1500-2000	14
2000-2500	56
2500-3000	60
3000-3500	86
3500-4000	74
4000-4500	62
4500-5000	48

Find the median life time of a lamp.

6. 100 surnames were randomly picked up from a local telephone directory and the frequency distribution of the number of letters in the English alphabets in the surnames was obtained as follows:

Number of letters	1-4	4-7	7-10	10-13	13-16	16-19
Number of surnames	6	30	40	16	4	4

Determine the median number of letters in the surnames. Find the mean number of letters in the surnames? Also, find the modal size of the surnames.

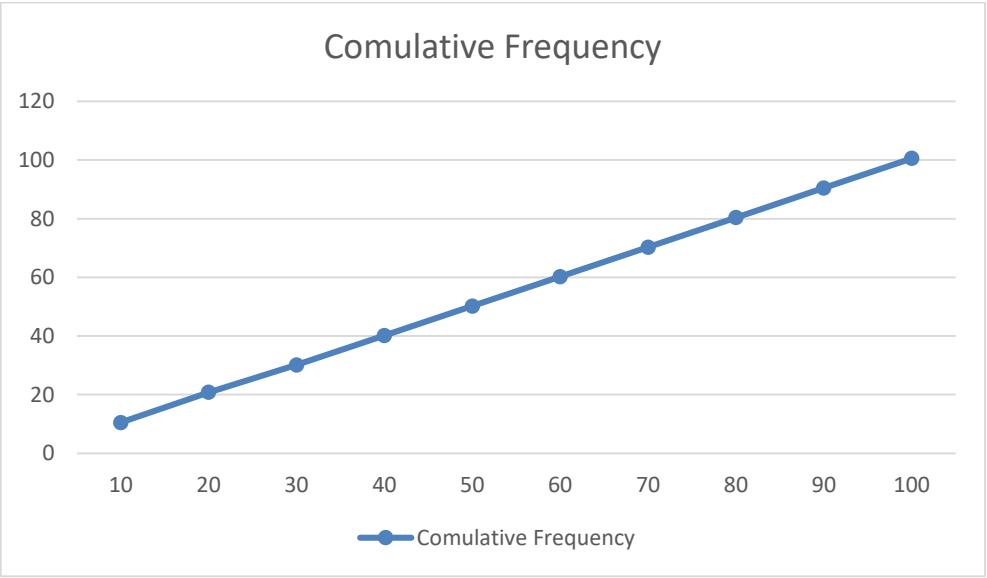
7. The distribution below gives the weights of 30 students of a class. Find the median weight of the students.

Weight (in kg)	40-45	45-50	50-55	55-60	60-65	65-70	70-75
Number of students	2	3	8	6	6	3	2

Graphical Representation of Cumulative Frequency Distribution

As we all know, pictures speak better than words. A graphical representation helps us in understanding given data at a glance. In chapter four, we have represented the data through bar graphs, histograms and frequency polygons. Let us now represent a cumulative frequency distribution graphically.

For example, let us consider the cumulative frequency distribution given in Table 5.13.

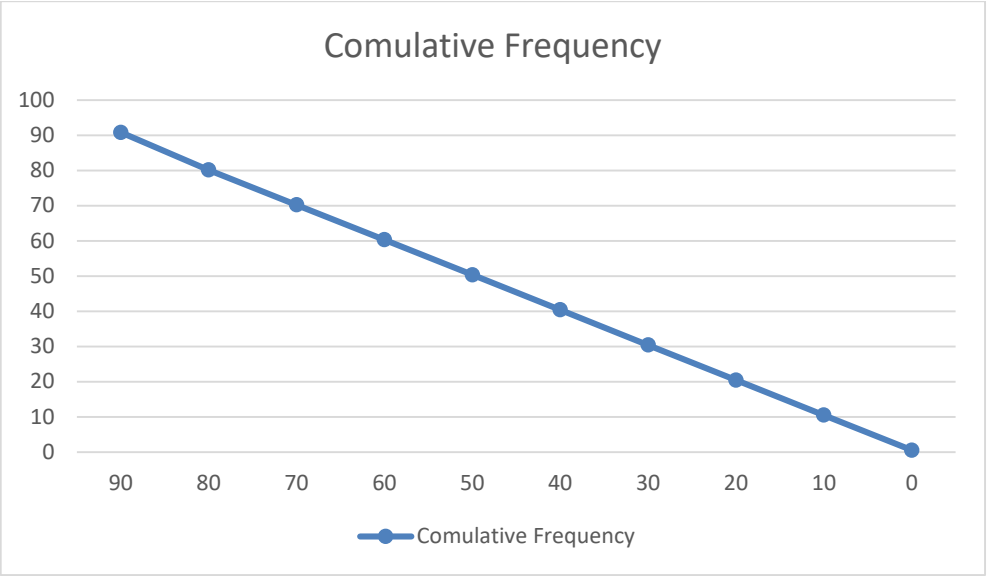


Recall that the value 10, 20, 30, …, 100 are the upper limits of the respective class intervals. To represent the data in the table graphically, we mark the upper limits of the class interval on the horizontal axis (x-axis) and their corresponding cumulative frequencies on the vertical axis (y-axis), choosing a convenient scale. The scale may not be the same on both the axis. Let us now plot the points corresponding to the ordered pairs given by (upper limit, corresponding cumulative frequency), i.e.,
$(10,5), (20,8), (30,12), (40,15),$
$(50,18), (60,22), (70,29), (80,38), (90,45), (100,53)$ on a graph paper and join them by a free hand smooth curve. The curve we get is called a cumulative frequency curve, or an ogive (of the less than type).

NOTE: The term 'Ogive' is pronounced as 'Ojeev' and is derived from the word 'Ogee'. An ogee is a shape consisting of a concave arc flowing into a convex arc, so forming an S-shaped curve with vertical ends. In architecture, the ogee shape in one of the characteristics of the 14^{th} and 15^{th} century Gothic styles.

Next, again we consider the cumulative frequency distribution given in Table 5.14 and draw its ogive (of the more than type).

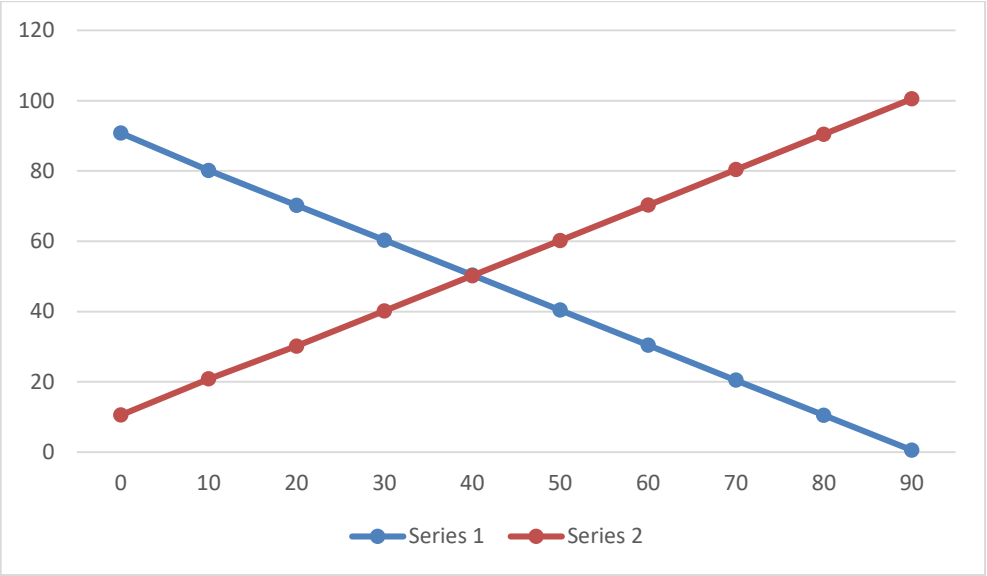
Recall that, here 0, 10, 20, …, 90 are the lower limits of the respective class intervals 0-10, 10-20, ..., 90-100. To represent 'the more than type' graphically, we plot the lower limits on the x- axis and the corresponding cumulative frequencies on the y-axis. Then we plot the points (lower limit, corresponding cumulative frequency), i.e., $(0,53), (10,48), (20,45), (30,41), (40,38), (50,35), (60,31), (70,24), (80,15), (90,8)$ on a graph paper, and join them by a free hand smooth curve.
The curve we get is a cumulative frequency curve, or an ogive (of the more than type).



Remark: Note that both the Ogives (in fig 5.1 and 5.2) corresponding to the same data, which is given in Table 5.12.

Now, are the ogives related to the median in any way? Is it possible to obtain the median from these two cumulative frequency curves corresponding to the data in Table 5.12? let us see.

One obvious way is to locate $\frac{n}{2} = \frac{53}{2} = 26.5$ on the y-axis (see fig 5.3). From this point, draw a line parallel to the x-axis cutting the curve at a point. From this point, draw a perpendicular to the x-axis. The point of intersection of this perpendicular with the x-axis determines the median of the data (see fig. 5.3).



Another way of obtaining the median is the following:
Draw both ogives (i.e., of the less than type and of the more than type) on the same axis. The two ogives will intersect each other at a point. From this point, if we draw a perpendicular on the x-axis, the point at which it cuts the x-axis gives us the median (see fig. 5.4).

Example 9: The annual profits earned by 30 shops of a shopping complex in a locality give rise to the following distribution:

Profit (in 100000 Afs)	Number of shops (Frequency)
$More\ than\ or\ equal\ to\ 5$	30
$More\ than\ or\ equal\ to\ 10$	28
$More\ than\ or\ equal\ to\ 15$	16
$More\ than\ or\ equal\ to\ 20$	14
$More\ than\ or\ equal\ to\ 25$	10
$More\ than\ or\ equal\ to\ 30$	7
$More\ than\ or\ equal\ to\ 35$	3

Draw both ogives for the data above. Hence obtain the median profit.
Solution: We first draw the coordinate axes, with lower limits of the profit along the horizontal axis, and the cumulative frequency along the vertical axes. Then, we plot the points $(5,30), (10,28), (15,16), (20,14), (25,10),$

$(30,7)$ *and* $(35,3)$. We join these points with a smooth curve to get the 'more than' ogive, as shown in fig 5.5.

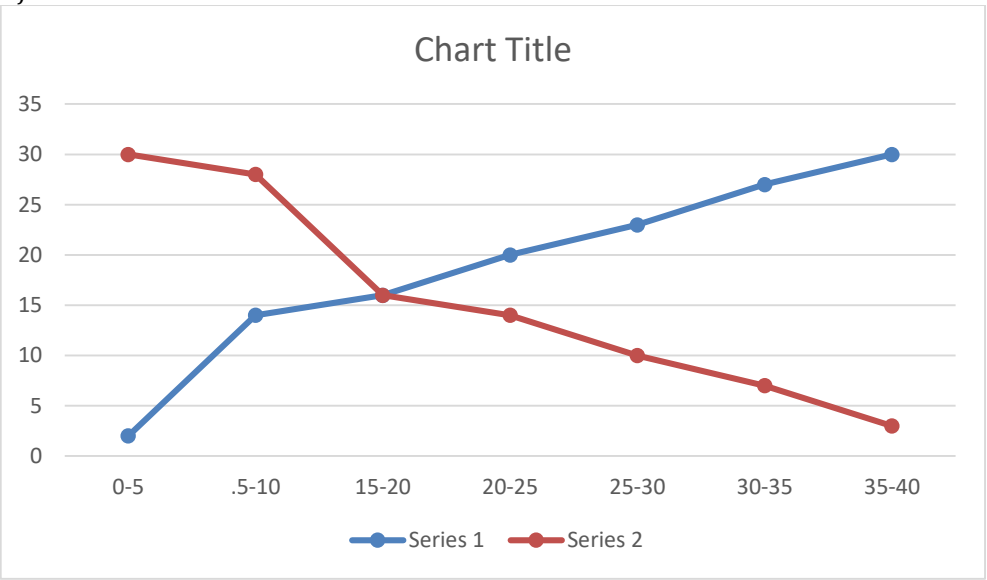
Now, let us obtain the classes, their frequencies and the cumulative frequency from the table above.

Table 5.17

Classes	5-10	10-15	15-20	20-25	25-30	30-35	35-40
No. of shops	2	12	2	4	3	4	3
Cumulative frequency	2	14	16	20	23	27	30

Using these values, we plot the points $(10,2)$, $(15,14)$, $(20,16)$, $(25,20)$, $(30,23)$, $(35,27)$, $(40,30)$ on the same axes as in fig 5.5 to get the 'less than' ogive, as shown in fig 5.6.

The abscissa of their point of intersection is nearly 17.5, which is the median. This can also be verified by using the formula. Hence, the median profit (in 100000 Afs) is $17.5 = Afs\ 1750000$.



Remark: In the above example, it may be noted that the class intervals were continuous. For drawing ogives, it should be ensured that the class intervals are continuous. (Also see constructions of histograms in chapter four).

Chapter Exercise

1. The following distribution gives the daily income of 50 workers of a factory.

Daily income (in Afs)	100-120	120-140	140-160	160-180	180-200
Number of workers	12	14	8	6	10

Convert the distribution above to a less than type cumulative frequency distribution, and draw its ogive.

2. During the medical check-up of 35 students of a class, their weights were recorded as follows:

Weight (in kg)	**Number of students**
$less\ than\ 38$	0
$less\ than\ 40$	3
$less\ than\ 42$	5
$less\ than\ 44$	9
$less\ than\ 46$	14
$less\ than\ 48$	28
$less\ than\ 50$	32
$less\ than\ 52$	35

Draw a less than type ogive for the given data. Hence obtain the median weight from the graph and verify the result by using the formula.

3. The following table gives production yield per hectare of wheat of 100 farms of a village.

Production yield (in kg/ha)	50-55	55-60	60-65	65-70	70-75	75-80
Number of farms	2	8	12	24	38	16

Change the distribution to a more than type distribution, and draw its ogive.

Chapter Summary

In this chapter, you have studied the following points:

1. The mean for grouped data can be found by:

$i.$ $\quad the\ direct\ method: \bar{x} = \dfrac{\sum f_i x_i}{\sum f_i},$

$ii.$ $\quad the\ assumed\ mean\ method: \bar{x} = a + \dfrac{\sum f_i x_i}{\sum f_i},$

$iii.$ $\quad the\ step\ deviation\ method: \bar{x} = a + \left(\dfrac{\sum f_i u_i}{\sum f_i}\right) \times h,$

With the assumption that the frequency of a class is centered at its mid-point, called its class mark.

2. The mode for grouped data can be found by using the formula:

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

Where symbols have their usual meanings.

3. The cumulative frequency of a class is the frequency obtained by adding the frequencies of all the classes preceding the given class.
4. The median for grouped data is formed by using the formula:

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

Where symbols have their usual meanings.

5. Representing a cumulative frequency distribution graphically as a cumulative frequency curve, or an ogive of the less than type and of the more than type.
6. The median of grouped data can be obtained graphically as the x-coordinate of the point of intersection of the two ogives for this data.

A NOTE TO THE READER

For calculating mode and median for grouped data, it should be ensured that the class intervals are continuous before applying the formulae. Same condition also apply for construction of an ogive. Further, in case of ogives, the scale may not be the same on both the axes.

7

Measures of Dispersion

Introduction

We know that statistics deals with data collected for specific purposes. We can make decisions about the data by analyzing and interpreting it. In earlier chapters, we have studied methods of representing data graphically and in tabular form. This representation reveals certain salient features or characteristics of data. We have also studied the methods of finding a representative value for the given data. This value is called the measure of central tendency. Recall mean (arithmetic mean), median and mode are three measures of central tendency. A measure of central tendency gives us a rough idea where data points are centered. But, in order to make better interpretation from the data, we should also have an idea how the data are scattered or how much they are bunched around a measure of central tendency.

Consider now the runs scored by two batsmen in their last ten matches as follows:

$Batsman\ A$: $30, 91, 0, 64, 42, 80, 30, 5, 117, 71$

$Batsman\ B$: $53, 46, 48, 50, 53, 53, 58, 60, 57, 52$

Clearly, the mean and median of the data are

	Batsman A	Batsman B
Mean	53	53
Median	53	53

Recall that, we calculate the mean of a data (denoted by $\bar{x}$) by dividing the sum of the observations by the number of observations, i.e.,

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

Also, the median is obtained by first arranging the data in ascending or descending order and applying the following rule.

If the number of observations is odd, then the median is $\left(\frac{n+1}{2}\right)^{th}$ observation.

If the number of observations is even, then median is the mean of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations.

We find that the mean and median of the runs scored by both the batsmen A and B are same i.e., 53. Can we say that the performance of two players is same? Clearly No, because the variability in the scores of batsman A is from 0 (minimum) to 117 (maximum). Whereas, the range of the runs scored by batsman B is from 46 to 60.

Thus, the measures of central tendency are not sufficient to give complete information about a given data. Variability is another factor which is required to be studied under statistics. Like 'measure of central tendency' we want to have a single number to describe variability. This single number is called a 'measure of dispersion'. In this chapter, we shall learn some of the important measures of dispersion and their methods of calculation for ungrouped and grouped data.

Measure of Dispersion:

The dispersion or scatter in a data is measured on the basis of the observations and the types of the measure of central tendency, used there. There are following measures of dispersion:

- Range,
- Quartile Deviation,
- Mean Deviation,
- Standard deviation.

In this chapter, we shall study all of these measures of dispersion except the quartile deviation.

Range:

Recall that, in the example of runs scored by two batsmen A and B, we had some idea of variability in the scores on the basis of minimum and maximum runs in each series. To obtain a single number for this, we find the difference of maximum and minimum values of each series. This difference is called the 'Range' of the data.

In case of batsman A, $Range = 117 - 0 = 117$ and for batsman B, $Range = 60 - 46 = 14$. Clearly, Range of A > Range of B. Therefore, the scores are scattered or dispersed in case of A while for B these are close to each other.

Thus, Range of a series = Maximum value – Minimum value.

The Range of data gives us a rough idea of variability or scatter but does not tell about the dispersion of the data from a measure of central tendency. For this purpose, we need some other measure of variability. Clearly, such measure must depend upon the difference (or deviation) of the values from the central tendency.

The important measures of dispersion, which depend upon the deviation of the observations from a central tendency are mean deviation and standard deviation. Let us discuss them in detail.

Mean Deviation:

Recall that the deviation of an observation x from a fixed value 'a' is the difference $x - a$. In order to find the dispersion of values of x from a central value a, we find the deviations about a. An absolute measure of dispersion is the mean of these deviations. To find the mean, we must obtain the sum of the deviations. But, we know that a measure of central tendency lies between the maximum and the minimum values of the set of observations. Therefore, some of the deviations will be negative and some positive. Thus, the sum of deviations may vanish. Moreover, the sum of the deviations from mean ($\bar{x}$) is zero.

Also

$$Mean\ of\ deviatios = \frac{sum\ of\ deviations}{Number\ of\ observations} = \frac{0}{n} = 0$$

Thus, finding the mean of deviations about mean is not of any use for us, as far as the measure of dispersion is concerned.

Remember that, in finding a suitable measure of dispersion, we require the distance of each value from a central tendency or a fixed number a. Recall, that the absolute value of the difference of two numbers gives the distance between the numbers when represented on a number line. Thus, to find the measure of dispersion from a fixed number a we may take the mean of the absolute values of the deviations from the central value. This mean is called the 'mean deviation'. Thus mean deviation about a central value a is the mean of the absolute values of the deviations of the observations from a. The mean deviation from a is denoted as M.D. (a). Therefore,

$$M.D.(a) = \frac{Sum\ of\ absolute\ values\ of\ deviations\ from\ a}{Number\ of\ observations}$$

Remark: Mean deviation may be obtained from any measure of central tendency. However, mean deviation from mean and median are commonly used in statistical studies.

Let us now learn how to calculate mean deviation about mean and mean deviation about median for various types of data

Mean Deviation for ungrouped data: Let n observations be $x_1, x_2, x_3, \cdots, x_n$. The following steps are involved in the calculation of mean deviation about mean or median:

Step 1: Calculate the measure of central tendency about which we are to find the mean deviation. Let it be a.

Step 2: Find the deviation of each x_i from a, i.e., $x_1 - a, x_2 - a, x_3 - a, \cdots, x_n - a$.

Step 3: Find the absolute values of the deviations, i.e., drop the minus sign $(-)$, if it is there, i.e., $|x_1 - a|, |x_2 - a|, |x_3 - a|, \cdots, |x_n - a|$.

Step 4: Find the mean of the absolute values of the deviations. This mean is the mean deviation about a, i.e.,

$$M.D.(a) = \frac{\sum_{i=1}^{n} |x_i - a|}{n}$$

Thus

$$M.D.(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|, where\ \bar{x} = Mean$$

And

$$M.D.(M) = \frac{1}{n}\sum_{i=1}^{n} |x_i - M|, where\ M = Median$$

Note: In this chapter, we shall use the symbol M to denote median unless stated otherwise. Let us now illustrate the steps of the above method in following examples.

Example 1: Find the mean deviation about the mean for the following data:

$$6, 7, 10, 12, 13, 4, 8, 12$$

Solution: We proceed step-wise and get the following:

Step 1: Mean of the given data is

$$\bar{x} = \frac{6 + 7 + 10 + 12 + 13 + 4 + 8 + 12}{8} = \frac{72}{8} = 9$$

Step 2: The deviations of the respective observations from the mean $\bar{x}$, i.e., $x_i - \bar{x}$ are

$$6 - 9,\ \ 7 - 9,\ \ 10 - 9,\ \ 12 - 9,\ \ 13 - 9,\ \ 4 - 9,\ \ 8 - 9,\ \ 12 - 9$$

Or

$$-3,\ \ -2,\ \ 1,\ \ 3,\ \ 4,\ \ -5,\ \ -1,\ \ 3$$

Step 3: The absolute values of the deviations, i.e., $|x_i - \bar{x}|$ are

$$3,\ \ 2,\ \ 1,\ \ 3,\ \ 4,\ \ 5,\ \ 1,\ \ 3$$

Step 4: The required mean deviation about the mean is

$$M.D.(\bar{x}) = \frac{\sum_{i=1}^{8} |x_i - \bar{x}|}{8} = \frac{3 + 2 + 1 + 3 + 4 + 5 + 1 + 3}{8} = 2.75$$

Note: Instead of carrying out the steps every time, we can carry on calculation, step-wise without referring to steps.

Example 2: Find the mean deviation about the mean for the following data:

$$12, 3, 18, 17, 4, 9, 17, 19, 20, 15, 8, 17, 2, 3, 16, 11, 3, 1, 0, 5$$

Solution: We have to first find the mean $(\bar{x})$ of the given data

$$\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = \frac{200}{20} = 10$$

The respective absolute values of the deviations from mean, i.e., $|x_i - \bar{x}|$ are

$$2, 7, 8, 7, 6, 1, 7, 9, 10, 5, 2, 7, 8, 7, 6, 1, 7, 9, 10, 5$$

Therefore

$$\sum_{i=1}^{20} |x_i - \bar{x}| = 124$$

And

$$M.D.(\bar{x}) = \frac{124}{20} = 6.2$$

Example 3: Find the mean deviation about the median for the following data:

$$3, 9, 5, 3, 12, 10, 18, 4, 7, 19, 21$$

Solution: Here the number of observations is 11 which is odd. Arranging the data into ascending order, we have:

$$3, 3, 4, 5, 7, 9, 10, 12, 18, 19, 21$$

Now

$$Median = \left(\frac{11+1}{2}\right)^{th} or\ 6^{th} observation = 9$$

The absolute values of the respective deviations from the median, i.e., $|x_i - M|$ are:

$$6, 6, 5, 4, 2, 0, 1, 3, 9, 10, 12$$

Therefore

$$\sum_{i=1}^{11} |x_i - M| = 58$$

And

$$M.D.(M) = \frac{1}{11}\sum_{i=1}^{11} |x_i - M| = \frac{1}{11} \times 58 = 5.27$$

Mean Deviation for Grouped Data

: We know that data can be grouped into two ways:

a. Discrete frequency distribution,
b. Continuous frequency distribution,

Let us discuss the method of finding mean deviation for both types of the data.

Discrete frequency distribution:

Let the given data consist of n distinct values $x_1, x_2, x_3, \cdots, x_n$ occurring with frequencies $f_1, f_2, f_3, \cdots, f_n$ respectively. This data can be represented in the tabular form as given below, and is called discrete frequency distribution:

$$\begin{array}{l} x : x_1\ x_2\ x_3\ \cdots x_n \\ f\ :\ f_1\ f_2\ f_3\ \cdots f_n \end{array}$$

- Mean deviation about mean

First of all we find the mean $\bar{x}$ of the given data by using the formula

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

Where $\sum_{i=1}^{n} f_i x_i$ denotes the sum of the products of observations x_i with their respective frequencies f_i and $N = \sum_{i=1}^{n} f_i$ is the sum of the frequencies.

Then, we find the deviations of observations x_i from the mean $\bar{x}$ and take their absolute values, i.e., $|x_i - \bar{x}|\ for\ all\ i = 1, 2, \cdots, n.$

After this, find the mean of the absolute values of the deviations, which is the required mean deviation about the mean. Thus

$$M.D.(\bar{x}) = \frac{\sum_{i=1}^{n} f_i\,|x_i - \bar{x}|}{\sum_{i=1}^{n} f_i} = \frac{1}{N}\sum_{i=1}^{n} f_i\,|x_i - \bar{x}|$$

- Mean deviation about median

To find mean deviation about median, we find the median of the given discrete frequency distribution. For this the observations are arranged in ascending order. After this the cumulative frequencies are obtained. Then, we identify the observation whose cumulative frequency is equal to or just greater than $\frac{N}{2}$, where N is the sum of frequencies. This value of observation lies in the middle of the data, therefore, it is the required median. After finding median, we obtain the mean of the absolute values of the deviations.

$$M.D.(M) = \frac{1}{N}\sum_{i=1}^{n} f_i\,|x_i - M|$$

Example 4: Find mean deviation about the mean for the following data:

x_i	2	5	6	8	10	12
f_i	2	8	10	7	8	5

Solution: let us make a table 6.1 of the given data and append other columns after calculations.

x_i	f_i	$f_i x_i$	$\lvert x_1 - \bar{x}\rvert$	$f_i \lvert x_i - \bar{x}\rvert$
2	2	4	5.5	11
5	8	40	2.5	20
6	10	60	1.5	15
8	7	56	0.5	3.5
10	8	80	2.5	20
12	5	60	4.5	22.5
	40	**300**		**92**

$$N = \sum_{i=1}^{6} f_i = 40, \qquad \sum_{i=1}^{6} f_i x_i = 300, \qquad \sum_{i=1}^{6} f_i |x_i - \bar{x}| = 92$$

Therefore

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{6} f_i x_i = \frac{1}{40} \times 300 = 7.5$$

And

$$M.D.(\bar{x}) = \frac{1}{N}\sum_{i=1}^{6} f_i |x_i - \bar{x}| = \frac{1}{40} \times 92 = 2.3$$

Example 5: Find the mean deviation about the median for the following data:

x_i	3	6	9	12	13	15	21	22
f_i	3	4	5	2	4	5	4	3

Solution: The given observations are already in ascending order. Adding a row corresponding to cumulative frequencies to the given data, we get (Table 6.2)

Table 6.2

x_i	3	6	9	12	13	15	21	22
f_i	3	4	5	2	4	5	4	3
$c.f.$	3	7	12	14	18	23	27	30

Now, $N = 30$ which is even.

Median is the mean of the 15th and 16th observations. Both of these observations lie in the cumulative frequency 18, for which the corresponding observation is 13.

Therefore,

$$Median\ M = \frac{15^{th}\ observation + \ 16^{th}\ observation}{2} = \frac{13 + 13}{2} = 13$$

Now, absolute value of the deviations from median, i.e., $|x_i - M|$ are shown in Table 6.3.

Table 6.3

$\lvert x_i - M \rvert$	10	7	4	1	0	2	8	9
f_i	3	4	5	2	4	5	4	3
$f_i \lvert x_i - M \rvert$	30	28	20	2	0	10	32	27

We have

$$\sum_{i=1}^{8} f_i = 30 \ \ and \ \ \sum_{i=1}^{8} f_i |x_i - M| = 149$$

Therefore

$$M.D.(M) = \frac{1}{N}\sum_{i=1}^{8} f_i |x_i - M| = \frac{1}{30} \times 149 = 4.97.$$

Continuous frequency distribution: A continuous frequency distribution is a series in which the data are classified into different class-intervals without gaps along with their respective frequencies.

For example, marks obtained by 100 students are presented in a continuous frequency distribution as follows:

Marks obtained	0-10	10-20	20-30	30-40	40-50	50-60
Number of students	12	18	27	20	17	6

Mean deviation about mean

While calculating the mean of a continuous frequency distribution, we had made the assumption that the frequency in each class is centered at its mid-points. Here also, we write the mid-points of each given class and proceed further as for a discrete frequency distribution to find the mean deviation.

Let us take the following example.

Example 6: Find the mean deviation about the mean for the following data.

Marks obtained	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Number of students	2	3	8	14	8	3	2

Solution: We make the following Table 6.4 from the given data:

Marks obtained	**Number of students**	**Mid-points**	$f_i x_i$	$\lvert x_i - \bar{x}\rvert$	$f_i \lvert x_i - \bar{x}\rvert$
10-20	2	15	30	30	60
20-30	3	25	75	20	60
30-40	8	35	280	10	80
40-50	14	45	630	0	0
50-60	8	55	440	10	80
60-70	3	65	195	20	60
70-80	2	75	150	30	60
	40		**1800**		**400**

Here

$$N = \sum_{i=1}^{7} f_i = 40, \qquad \sum_{i=1}^{7} f_i x_i = 1800, \qquad \sum_{i=1}^{7} f_i \lvert x_i - \bar{x}\rvert = 400$$

Therefore

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{7} f_i x_i = \frac{1800}{40} = 45$$

And

$$M.D.(\bar{x}) = \frac{1}{N}\sum_{i=1}^{7} f_i \lvert x_i - \bar{x}\rvert = \frac{1}{40} \times 400 = 10$$

Shortcut method for calculating mean deviation about mean: We can avoid the tedious calculations of computing $\bar{x}$ by following step-deviation method. Recall that in this method, we take an assumed mean which is in the middle or just close to it in

the data. Then deviations of observations (or mid-point of classes) are taken from the assumed mean. This is nothing but the shifting of origin from zero to the assumed mean on the number line, as shown bellow:

After deviations

$$-60, -50, -40, -30, -20, -10, 0, 10, 20, 30, 40, 50, 60$$

Before deviations

$$0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120$$

$$Assumed\ Mean = 60$$

If there is a common factor of all the deviations, we divide them by this common factor to further simplify the deviations. These are known as step-deviations. The process of taking step-deviations is the change of scale on the number line as shown below:

Step deviation

$$-6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6$$

Deviation from

$$-60, -50, -40, -30, -20, -10, 0, 10, 20, 30, 40, 50, 60$$

Assumed mean

$$0, 10, 20, 30, 40, 50, \{60 = Assumed\ mean\}, 70, 80, 90, 100, 110, 120$$

The deviations and step-deviations reduce the size of the observations, so that the computations viz. multiplication, etc., become simpler. Let, the new variable be denoted by $d_i = \frac{x_i - a}{h}$, where a is the assumed mean and h is the common factor. Then, the mean $\bar{x}$ by step-deviation method is given by

$$\bar{x} = a + \frac{\sum_{i=1}^{n} f_i d_i}{N} \times h$$

Let us take the data of example 6 and find the mean deviation by using step-deviation method.

Take the assumed mean $a = 45\ and\ h = 10$, and form the following Table 6.5.

Table 6.5

Marks obtained	**Number of students (f_i)**	**Mid-points (x_i)**	$\boldsymbol{d_i = \frac{x_i - 45}{10}}$	$\boldsymbol{f_i d_i}$	$\boldsymbol{\lvert x_i - \bar{x} \rvert}$	$\boldsymbol{f_i \lvert x_i - \bar{x} \rvert}$
10-20	2	15	-3	-6	30	60
20-30	3	25	-2	-6	20	60
30-40	8	35	-1	-8	10	80
40-50	14	45	0	0	0	0
50-60	8	55	1	8	10	80
60-70	3	65	2	6	20	60
70-80	2	75	3	6	30	60
Total	**40**			**0**		**400**

Therefore

$$\bar{x} = a + \frac{\sum_{i=1}^{7} f_i d_i}{N} \times h = 45 + \frac{0}{40} \times 10 = 45$$

And

$$M.D.(\bar{x}) = \frac{1}{N}\sum_{i=1}^{7} f_i \, |x_i - \bar{x}| = \frac{400}{40} = 10$$

Note: The step deviation method is applied to compute $\bar{x}$. Rest of the procedure is the same.

Mean deviation about median

The process of finding the mean deviation about median for a continuous frequency distribution is similar as we did for mean deviation about the mean. The only difference lies in the replacement of the mean by median while talking deviations.

Let us recall the process of finding median for a continuous frequency distribution.

The data is first arranged in ascending order. Then, the median of continuous frequency distribution is obtained by first identifying the class in which median lies (median class) and then applying the formula

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

Where median class is the class interval whose cumulative frequency is just greater than or equal to $\frac{N}{2}$, N is the sum of frequencies, l, f, h and C are, respectively the lower limit, the frequency, the width of the median class and C the cumulative frequency of the class just preceding the median class. After finding the median, the absolute values of the deviations of mid-point x_i of each class from the median i.e., $|x_i - M|$ are obtained. Then

$$M.D.(M) = \frac{1}{N}\sum_{i=1}^{n} f_i \, |x_i - M|$$

The process is illustrated in the following example:

Example 7: Calculate the mean deviation about median for the following data:

Class	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	6	7	15	16	4	2

Solution: From the following Table 6.6. from the given data:

Table 6.6.

Class	**Frequency** (f_i)	**Cumulative Frequency** $(c.f)$	**Mid-Points** (x_i)	$\|x_i - Med\|$	$f_i\|x_i - Med\|$
0-10	6	6	5	23	138
10-20	7	13	15	13	91
20-30	15	28	25	3	45
30-40	16	44	35	7	112
40-50	4	48	45	17	68
50-60	2	50	55	27	54

	50				**508**

The class interval containing $\left(\frac{N}{2}\right)^{th}$ or 25th items is 20-30. Therefore, 20-30 is the median class. We know that

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

Here $l = 20, C = 13, f = 15, h = 10$ and $N = 50$, Therefore,

$$Median = 20 + \frac{25 - 13}{15} \times 10 = 20 + 8 = 28$$

Thus, mean deviation about median is given by

$$M.D.(M) = \frac{1}{N}\sum_{i=1}^{n} f_i\,|x_i - M| = \frac{1}{50} \times 508 = 10.16$$

EXERCISE 6.1

Find the mean deviation about the mean for the data in Exercise 1 and 2.

1. $4, 7, 8, 9, 10, 12, 13, 17.$
2. $38, 70, 48, 40, 42, 55, 63, 46, 54, 44.$

Find the mean deviation about the median for the data in Exercise 3 and 4.

3. $13, 17, 16, 14, 11, 13, 10, 16, 11, 18, 12, 17.$
4. $36, 72, 46, 42, 60, 45, 53, 46, 51, 49.$

Find the mean deviation about the mean for the data in Exercise 5 and 6.

5. $\begin{cases} x_i & 5, & 10, & 15, & 20, & 25 \\ f_i & 7, & 4, & 6, & 3, & 5 \end{cases}$
6. $\begin{cases} x_i & 10, & 30, & 50, & 70, & 90 \\ f_i & 4, & 24, & 28, & 16, & 8 \end{cases}$

Find the mean deviation about the median for the data in Exercise 7 and 8.

7. $\begin{cases} x_i & 5, & 7, & 9, & 10, & 12, & 15 \\ f_i & 8, & 6, & 2, & 2, & 2, & 6 \end{cases}$
8. $\begin{cases} x_i & 15, & 21, & 27, & 30, & 35, \\ f_i & 3, & 5, & 6, & 7, & 8 \end{cases}$

Find the mean deviation about the mean for the data in Exercises 9 and 10.

9. Income $0-100, 100-200, 200-300, 300-400, 400-500, 500-600, 600-700, 700-800$ per day
Number $4\ \ 8\ \ 9\ \ 10\ \ 7\ \ 5\ \ 4\ \ 3$ of persons.

10. height $95-105,\ 105-115,\ 115-125,\ 125-135,\ 135-145,\ 145-155$ in cm.
$Number\ of\ 9\ \ 13\ \ 26\ \ 30\ \ 12\ \ 10\ \ boys.$

11. Find the mean deviation about median for the following data:
Marks 0-10, 20-20, 20-30, 30-40, 40-50, 50-60,
Number of students 6 8 14 16 4 2

12. Calculate the mean deviation about median age for the age distribution of 100 persons given below:

Age 16-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55
Number 5 6 12 14 26 12 16 9
{Hint: Convert the given data into continuous frequency distribution by subtracting 0.5 from the lower limit and adding 0.5 to the upper limit of each class interval}

Limitations of mean deviation: In a series, where the degree of variability is very high, the median is not a representative central tendency. Thus, the mean deviation about median calculated for such series cannot be fully relied.

The sum of the deviations from the mean (minus signs ignored) is more than the sum of the deviations from median. Therefore, the mean deviation about the mean is not very scientific. Thus, in many cases, mean deviation may give unsatisfactory results. Also mean deviation is calculated on the basis of absolute values of the deviations and therefore, cannot be subjected to further algebraic treatment. This implies that we must have some other measure of dispersion. Standard deviation is such a measure of dispersion.

Variance and Standard Deviation:

Recall that while calculating mean deviation about mean or median, the absolute values of the deviations were taken. The absolute values were taken to give meaning to the mean deviation, otherwise the deviations may cancel among themselves.

Another way to overcome this difficulty which arose due to the signs of deviations, is to take squares of all the deviations. Obviously all these squares of deviations are non-negative. Let $x_1, x_2, x_3, \cdots, x_n$ be n observations and $\bar{x}$ be their mean. Then

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 +, \cdots, +(x_n - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

If this sum is zero, then each $(x_i - \bar{x})$ has to be zero. This implies that there is no dispersion at all as all observations are equal to the mean x.

If $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is small, this indicates that the observations $x_1, x_2, x_3, \cdots, x_n$ are close to the mean $\bar{x}$ and therefore, there is a lower degree of dispersion. On the contrary, if this sum is large, there is a higher degree of dispersion of the observations from the mean $\overline{x}$. Can we thus say that the sum $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is a reasonable indicator of the degree of dispersion or scatter?

Let us take the set A of six observations $5, 15, 25, 35, 45, 55$. The mean of the observations is $\bar{x} = 30$. The sum of squares of deviations from $\bar{x}$ for this set is

$$\sum_{i=1}^{6}(x_i - \bar{x})^2 = (5-30)^2 + (15-30)^2 + (25-30)^2 + (35-30)^2 + (45-30)^2 + (55-30)^2 = 625 + 225 + 25 + 25 + 225 + 625 = 1750$$

Let us now take another set B of 31 observations
$15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35,$

$36, 37, 38, 39, 40, 41, 42, 43, 44, 45$. The mean of these observations is $\bar{y} = 30$.

Note that both the sets A and B of observations have a mean of 30.

Now, the sum of squares of deviations of observations for set B from the mean y is given by

$$\sum_{i=1}^{31}(y_i - \bar{y})^2 = (15-30)^2 + (16-30)^2 + (17-30)^2 +, \cdots, +(44-30)^2 + (45-30)^2$$

$$= (-15)^2 + (-14)^2 +, \cdots, +(-1)^2 + 0^2 + 1^2 + 2^2 + 3^2 +, \cdots, +14^2$$

$$+ 15^2 = 2[15^2 + 14^2 +, \cdots, +1^2] = 2 \times \frac{15 \times (15+1)(30+1)}{6}$$

$$= 5 \times 16 \times 31 = 2480$$

(Because sum of squares of first n natural numbers $= \frac{n(n+1)(2n+1)}{6}$. Here $n = 15$)

If $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is simply our measure of dispersion or scatter about mean, we will tend to say that the set A of six observations has a lesser dispersion about the mean than the set B of 31 observations, even though the observations in set A are more scattered from the mean (the range of deviations being from -25 to 25) than in the set B (where the range of deviations is from -15 to 15).

Thus, we can say that the sum of squares of deviations from the mean is not a proper measure of dispersion. To overcome this difficulty we take the mean of the squares of the deviations, i.e., we take $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. In case of the set A, we have $Mean = \frac{1}{6} \times 1750 = 291.67$ and in case of the set B, it is $\frac{1}{31} \times 2480 = 80$.

This indicates that the scatter or dispersion is more in set A than the scatter or dispersion in set B, which confirms with the geometrical representation of the two sets.

Thus, we can take $\frac{1}{n}\sum(x_i - \bar{x})^2$ as a quantity which leads to a proper measure of dispersion. This number, i.e., mean of the squares of the deviations from mean is called the variance and is denoted by σ^2(read sigma square). Therefore, the variance of n observations $x_1, x_2, x_3, \cdots, x_n$ is given by

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard Deviation: In the calculation of variance, we find that the units of individual observations x_i and the unit of their mean $\bar{x}$ are different from that of variance, since variance involves the sum of squares of $(x_i - \bar{x})$. For this reason, the proper measure of dispersion about the mean of a set of observations is expressed as positive square root of the variance and is called standard deviation. Therefore, the standard deviation, usually denoted by σ, is given by

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (I)$$

Let us take the following example to illustrate the calculation of variance and hence, standard deviation of ungrouped data.

Example 8: Find the variance of the following data:

$$6, 8, 10, 12, 14, 16, 18, 20, 22, 24$$

Solution: From the given data we can form the following Table 6.7. The mean is calculated by step-deviation method taking 14 as assumed mean. The number of observations is $n = 10$

Table 6.7.

x_i	$d_i = \frac{x_i - 14}{2}$	$Deviations\ from\ mean$ $(x_i - \bar{x})$	$(x_i - \bar{x})$
6	-4	-9	81
8	-3	-7	49
10	-2	-5	25
12	-1	-3	9
14	0	-1	1
16	1	1	1
18	2	3	9
20	3	5	25
22	4	7	49
24	5	9	81
	5		330

Therefore, Mean $\bar{x} = assumed\ mean + \frac{\sum_{i=1}^{n} d_i}{n} \times h = 14 + \frac{5}{10} \times 2 = 15$ and Variance

$$(\sigma^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10} \times 330 = 33$$

Thus standard deviation $(\sigma) = \sqrt{33} = 5.74$.

Standard Deviation of a discrete frequency distribution: Let the given discrete frequency distribution be

$$\begin{array}{ll} x: & x_1, \quad x_2, \quad x_3, \cdots, x_n \\ f: & f_1, \quad f_2, \quad f_3, \cdots, f_n \end{array}$$

In this case standard deviation

$$(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \bar{x})^2} \qquad (II)$$

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

Let us take up following example.

Example 9: Find the variance and standard deviation for the following data:

x_i	4	8	11	17	20	24	32

f_i	3	5	9	5	4	3	1

Solution: Presenting the data in tabular form (Table 6.8), we get

Table 6.8.

x_i	f_i	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
4	3	12	-10	100	300
8	5	40	-6	36	180
11	9	99	-3	9	81
17	5	85	3	9	45
20	4	80	6	36	144
24	3	72	10	100	300
32	1	32	18	324	324
	30	**420**			**1374**

$$N = 30, \qquad \sum_{i=1}^{7} f_i x_i = 420, \qquad \sum_{i=1}^{7} f_i (x_i - \bar{x})^2 = 1374$$

Therefore

$$\bar{x} = \frac{\sum_{i=1}^{7} f_i x_i}{N} = \frac{1}{30} \times 420 = 14$$

Hence

$$variance\ (\sigma^2) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{30} \times 1374 = 45.8$$

And

$$standard\ deviation\ (\sigma) = \sqrt{\frac{1}{N} \sum_{i=1}^{7} f_i\ (x_i - \bar{x})^2} = \sqrt{45.8} = 6.77$$

Standard Deviation of a continuous frequency distribution: The given continuous frequency distribution can be represented as a discrete frequency distribution by replacing each class by its mid-points. Then, the standard deviation is calculated by the technique adopted in the case of a discrete frequency distribution.

If there is a frequency distribution of n classes each class defined by its mid-point x_i with frequency f_i, the standard deviation will be obtained by the formula

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

Where $\bar{x}$ is the mean of the distribution and $N = \sum_{i=1}^{n} f_i$.

Another formula for standard deviation: We know that

$$Variance\ (\sigma^2) = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i^2 + (\bar{x})^2 - 2(\bar{x})x_i)$$

$$= \frac{1}{N}\left[\sum_{i=1}^{n} f_i x_i^2 + \sum_{i=1}^{n} f_i(\bar{x})^2 - \sum_{i=1}^{n} 2(\bar{x}) f_i x_i\right]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{n} f_i x_i^2 + (\bar{x})^2 \sum_{i=1}^{n} f_i - 2(\bar{x})\sum_{i=1}^{n} f_i x_i\right]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{n} f_i x_i^2 + (\bar{x})^2\, N - 2(\bar{x})N(\bar{x})\right]\left[Here\ \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2\right.$$

$$\left.= (\bar{x})\ or\ \sum_{i=1}^{n} x_i f_i = N(\bar{x})\right] = \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i^2 + (\bar{x})^2 - 2(\bar{x})^2$$

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

Or

$$(\sigma^2) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - \left(\frac{\sum_{i=1}^{n} f_i x_i}{N}\right)^2 = \frac{1}{N^2}\left[N\sum_{i=1}^{n} f_i x_i^2 - \left(\sum_{i=1}^{n} f_i x_i\right)^2\right]$$

Thus, standard deviation

$$(\sigma) = \frac{1}{N}\sqrt{N\sum_{i=1}^{n} f_i\, x_i^2 - \left(\sum_{i=1}^{n} f_i x_i\right)^2} \qquad (III)$$

Example 10: Calculate the mean, variance and standard deviation for the following distribution:

Class

$$30-40\ \ 40-50\ \ 50-60\ \ 60-70\ \ 70-80\ \ 80-90\ \ 90-100$$

Frequency

$$3\ \ 7\ \ 12\ \ 15\ \ 8\ \ 3\ \ 2$$

Solution: From the given data, we construct the following Table 6.9.

Table 6.9.

Class	**Frequency** (f_i)	**Mid-point** (x_i)	$f_i x_i$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
30-40	3	35	105	729	2187
40-50	7	45	315	289	2023
50-60	12	55	660	49	588
60-70	15	65	975	9	135
70-80	8	75	600	169	1352
80-90	3	85	255	529	1587

90-100	2	95	190	1089	2178
	50		**3100**		**10050**

Thus

$$Mean\ \bar{x} = \frac{1}{N}\sum_{i=1}^{7} f_i x_i = \frac{3100}{50} = 62$$

$$Variance\ (\sigma^2) = \frac{1}{N}\sum_{i=1}^{7} f_i\,(x_i - x)^2 = \frac{1}{50} \times 10050 = 201$$

And

$$standard\ deviation\ (\sigma) = \sqrt{201} = 14.18$$

Example 11: Find the standard deviation for the following data:

x_i	3	8	13	18	23
f_i	7	10	15	10	6

Solution: Let us form the following Table 6.10:

Table 6.10.

x_i	f_i	$f_i x_i$	x_i^2	$f_i x_i^2$
3	7	21	9	63
8	10	80	64	640
13	15	195	169	2535
18	10	180	324	3240
23	6	138	529	3174
		614		**9652**

Now, by formula (III), we have

$$\sigma = \frac{1}{N}\sqrt{N\sum f_i x_i^2 - \left(\sum f_i x_i\right)^2} = \frac{1}{48}\sqrt{48 \times 9652 - (614)^2} = \frac{1}{48}\sqrt{463296 - 376996}$$

$$= \frac{1}{48} \times 293.77 = 6.12$$

Therefore, Standard deviation $(\sigma) = 6.12$.

Shortcut method to find variance and standard deviation: Sometimes the values of x_i in a discrete distribution or the mid points x_i of different classes in a continuous distribution are large and so the calculation of mean and variance becomes tedious and time consuming. By using step-deviation method, it is possible to simplify the procedure.

Let the assumed mean be 'A' and the scale be reduced to $\frac{1}{h}$ times (h being the width of class-intervals). Let the step- deviation or the new values be y_i.

$$y_i = \frac{x_i - A}{h} \quad or \quad x_i = A + hy_i \qquad (I)$$

We know that

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N} \qquad (II)$$

Replacing x_i from (I) in (II), we get

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i(A + hy_i)}{N} = \frac{1}{N}\left(\sum_{i=1}^{n} f_i A + \sum_{i=1}^{n} hf_i y_i\right) = \frac{1}{N}\left(A\sum_{i=1}^{n} f_i + h\sum_{i=1}^{n} f_i y_i\right)$$

$$= A\frac{N}{N} + h\frac{\sum_{i=1}^{n} f_i y_i}{N} \qquad \left(\because \sum_{i=1}^{n} f_i = N\right)$$

$$\bar{x} = A + h\bar{y} \qquad (III)$$

Variance of the variable x,

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x}) = \frac{1}{N}\sum_{i=1}^{n} f_i\,(A + hy_i - A - h\bar{y})^2 = \frac{1}{N}\sum_{i=1}^{n} f_i\, h^2(y_i - \bar{y})^2$$

$$= \frac{h^2}{N}\sum_{i=1}^{n} f_i(y_i - \bar{y})^2 = h^2 \times variance\ of\ the\ variable\ y_i$$

i.e.,

$$\sigma_x^2 = h^2\sigma_y^2$$

Or

$$\sigma_x = h\sigma_x \qquad (IV)$$

From $(III)\ and\ (IV)$, we have

$$\sigma_x = \frac{h}{N}\sqrt{N\sum_{i=1}^{n} f_i\, y_i^2 - \left(\sum_{i=1}^{n} f_i y_i\right)^2} \qquad (V)$$

Let us solve Example 11 by the short-cut method and using formula (V).

Example 12: calculate mean, variance and standard deviation for the following distribution.

Classes	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	3	7	12	15	8	3	2

Solution: Let the assumed mean $A = 65$. Here $h = 10$. We obtain the following Table 6.11 from the given data:

Class	Frequency (f_i)	Mid-point (x_i)	$y_i = \frac{x_i - 65}{10}$	y_i^2	$f_i y_i$	$f_i y_i^2$
30-40	3	35	-3	9	-9	27
40-50	7	45	-2	4	-14	28
50-60	12	55	-1	1	-12	12
60-70	15	65	0	0	0	0
70-80	8	75	1	1	8	8
80-90	3	85	2	4	6	12
90-100	2	95	3	9	6	18
	$N = 50$				**-15**	**105**

Therefore,

$$\bar{x} = A + \frac{\sum f_i y_i}{50} \times h = 65 - \frac{15}{50} \times 10 = 62$$

Variance

$$\sigma^2 = \frac{h^2}{N^2}\left[N\sum f_i y_i^2 - \left(\sum f_i y_i\right)^2\right] = \frac{(10)^2}{(50)^2}[50 \times 105 - (-15)^2] = \frac{1}{25}[5250 - 225]$$
$$= 201$$

And standard deviation $(\sigma) = \sqrt{201} = 14.18$

EXERCISE

Find the mean and variance for each of the data in exercise 1 to 5.

1. $6, 7, 10, 12, 13, 4, 8, 12$
2. $First\ n\ natural\ numbers$
3. $first\ 10\ multiples\ of\ 3$
4.

x_i	6	10	14	18	24	28	30
f_i	2	4	7	12	8	4	3

5.

x_i	92	93	97	98	102	104	109
f_i	3	2	3	2	6	3	3

6. $Find\ the\ mean\ and\ standard\ deviation\ using\ short - cut\ method$

x_i	60	61	62	63	64	65	66	67	68
f_i	2	1	12	29	25	12	10	4	5

Find the mean and variance for the following frequency distributions in exercise 7 and 8.

7.

Classes	0-30	30-60	60-90	90-120	120-150	150-180	180-210
Frequency	2	3	5	10	3	5	2

8.

Classes	0-10	10-20	20-30	30-40	40-50
Frequency	5	8	15	16	6

9. *Find the mean, variance and standard deviation using short − cut method*

Height in cms	70-75	75-80	80-85	85-90	90-95	95-100	100-105	105-110	110-115
No. of children	3	4	7	7	15	9	6	6	3

10. *The diameters of circles (in mm)drawn in a design are given below*:

Diameters	33-36	37-40	41-44	45-48	49-52
No. of circles	15	17	21	22	25

Calculate the standard deviation and mean diameter of the classes.
[Hint: First make the data continuous by making classes as 32.5-36.5, 36.5-40.5… and then proceed.]

Analysis of Frequency Distributions:

In earlier sections, we have studied about some types of measures of dispersion. The mean deviation and the standard deviation have the same units in which the data are given. Whenever we want to compare the variability of two series with same mean, which are measured in different units, we do not merely calculate the measure of dispersion but we require such measures which are independent of the units. The measure of variability which is independent of units is called coefficient of variation (denoted as C.V.)

The coefficient of variation is defined as

$$C.V. = \frac{\sigma}{x} \times 100, \qquad \bar{x} \neq 0,$$

Where σ and $\bar{x}$ are the standard deviation and mean of the data.

For comparing the variability or dispersion of two series, we calculate the coefficient of variance for each series. The series having greater C.V. is said to be more variable than the other. The series having lesser C.V. is said to be more consistent than the other.

Comparison of two frequency distributions with same mean: Let $\overline{x_1}$ and σ_1 be the mean and standard deviation of the first distribution, and $\overline{x_2}$ and σ_2 be the mean and standard deviation of the second distribution.

Then

$$C.V.(1^{st}\ distribution) = \frac{\sigma_1}{\overline{x_1}} \times 100$$

And

$$C.V.(2^{nd}\ distribution) = \frac{\sigma_2}{\overline{x_2}} \times 100$$

$$\overline{x_1} = \overline{x_2} = \bar{x}\ \ (say)$$

Therefore

$$C.V.(1^{st}\ distribution) = \frac{\sigma_1}{\bar{x}} \times 100 \qquad (i)$$

And

$$C.V.(2^{nd}\ distribution) = \frac{\sigma_2}{x} \times 100 \qquad (ii)$$

It is clear from (i) and (ii) that the two C.Vs. can be compared on the basis of values of σ_1 and σ_2 only.

Thus, we say that for two series with equal means, the series with greater standard deviation (or variance) is called more variable or dispersed than the other. Also, the series with lesser value of standard deviation (or variance) is said to be more consistent than the other.

Let us now take following examples:

Example 13: Two plants A and B of a factory show following results about the number of workers and the wages paid to them.

	A	B
$Number\ of\ workers$	5000	6000
$Average\ monthly\ wages$	2500	2500
$Variance\ of\ distribution\ of\ wages$	81	100

In which plant, A or B is there greater variability in individual wages?

Solution: The variance of the distribution of wages in plant A $(\sigma_1^2) = 81$. Therefore, standard deviation of the distribution of wages in plant A $(\sigma_1) = 9$.

Also, the variance of the distribution of wages in plant B $(\sigma_2^2) = 100$. Therefore, standard deviation of the distribution of wages in plant B $(\sigma_1) = 10$. Since the average monthly wages in both the plants is same, i.e., 2500, therefore, the plant with greater standard deviation will have more variability? Thus, the plant B has greater variability in the individual wages.

Example 14: coefficient of variation of two distributions are 60 and 70, and their standard deviations are 21 and 16, respectively. What are their arithmetic means?

Solution: Given

$$C.V.(1^{st}\ distribution) = 60, \qquad \sigma_1 = 21$$

$$C.V.(2^{nd}\ distribution) = 70, \qquad \sigma_2 = 16$$

Let $\overline{x_1}$ and $\overline{x_2}$ be the means of 1st and 2nd distribution, respectively. Then

$$C.V.(1^{st}\ distribution) = \frac{\sigma_1}{\bar{x}} \times 100$$

Therefore

$$60 = \frac{21}{\overline{x_1}} \times 100 \quad or \quad \overline{x_1} = \frac{21}{60} \times 100 = 35$$

And

$$C.V.(2^{nd}\ distribution) = \frac{\sigma_2}{\overline{x_2}} \times 100$$

i.e.,

$$70 = \frac{16}{\overline{x_2}} \times 100 \quad or \quad \overline{x_2} = \frac{16}{70} \times 100 = 22.85$$

Example 15: The following values are calculated in respect of heights and weights of students of a section of class BBA.

	$Height$	$Weight$
$Mean$	$162.6\ cm$	$52.36\ kg$
$Variance$	$127.69\ cm^2$	$23.1361\ kg^2$

Can we say that the weights show greater variation than the heights?

Solution: To compare the variability, we have to calculate their coefficients of variation.

Given

$$Variance\ of\ heigth = 127.69\ cm^2$$

Therefore

$$Standard\ deviation\ of\ height = \sqrt{127.69}\ cm = 11.3\ cm$$

Also

$$Variance\ of\ weight = 23.1361\ kg^2$$

Therefore, standard deviation of weight $= \sqrt{23.1361}\ kg = 4.81\ kg.$

Now, the coefficient of variations (C.V.) are given by

$$(C.V.)\ in\ heights = \frac{Standard\ Deviation}{Mean} \times 100 = \frac{11.3}{162.6} \times 100 = 6.95$$

And

$$(C.V.)\ in\ weights = \frac{4.81}{52.36} \times 100 = 9.18$$

Clearly C.V. in weights is greater than the C.V. in heights. Therefore, we can say that weights show more variability than heights.

Chapter Exercise

1. From the data given below state which group is more variable, A or B?

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Group A	9	17	32	33	40	10	9
Group B	10	20	30	25	43	15	7

2. From the prices of shares X and Y below, find out which is more stable in value:

X	35	54	52	53	56	58	52	50	51	49
Y	108	107	105	105	106	107	104	103	104	101

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

	Firm A	*Firm B*
Number of wage earners	586	648
Mean of monthly wages	5253	5253
Variance of the distribution of wages	100	121

i. Which firm A or B pays larger amount as monthly wages?

ii. Which firm, A or B, shows greater variability in individual wages?

4. The following is the record of goals scored by team A in a football session:

Number of goals scored	0	1	2	3	4
Number of matches	1	9	7	5	3

For the team B, mean number of goals scored per match was 2 with a standard deviation 1.25 goals. Find which team may be considered more consistent?

5. The sum and sum of squares corresponding to length x (in cm) and weight y (in gm) of 50 plant products are given below:

$$\sum_{i=1}^{50} x_i = 212, \quad \sum_{i=1}^{50} x_i^2 = 902.8, \quad \sum_{i=1}^{50} y_i = 261, \quad \sum_{i=1}^{50} y_i^2 = 1457.6$$

Which is more varying, the length or weight?

Miscellaneous Examples

Example 16: The variance of 20 observations is 5. If each observation is multiplied by 2, find the new variance of the resulting observations.

Solution: Let the observations be $x_1, x_2, \cdots, x_{20}$ and $\bar{x}$ be their mean. Given that variance =5 and n=20. We know that

$$Variance\ (\sigma^2) = \frac{1}{n}\sum_{i=1}^{20}(x_i - \bar{x})^2, \qquad i.e., \quad 5 = \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^2$$

Or

$$\sum_{i=1}^{20}(x_i - \bar{x})^2 = 100 \qquad (i)$$

If each observation is multiplied by 2, and the new resulting observations are y_i, then

$$y_i = 2x_i \quad i.e., \qquad x_i = \frac{1}{2}y_i$$

Therefore

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{20} y_i = \frac{1}{20}\sum_{i=1}^{20} 2x_i = 2 \cdot \frac{1}{20}\sum_{i=1}^{20} x_i$$

i.e.,

$$\bar{y} = 2\bar{x} \quad or \quad \bar{x} = \frac{1}{2}\bar{y}$$

Substituting the values of x_i and $\bar{x}$ in (i), we get

$$\sum_{i=1}^{20}\left(\frac{1}{2}y_i - \frac{1}{2}\bar{y}\right)^2 = 100, \qquad i.e., \qquad \sum_{i=1}^{20}(y_i - \bar{y})^2 = 400$$

Thus the variance of new observations $= \frac{1}{20} \times 400 = 20 = 2^2 \times 5$.

Note: the reader may note that if each observation is multiplied by a constant k, the variance of the resulting observations becomes k^2 times the original variance.

Example 17: The mean of 5 observations in 4.4 and their variance is 8.24. if three of the observations are 1, 2 and 6, find the other two observations.

Solution: Let the other two observations be x and y.

Now

$$Mean\ \bar{x} = 4.4 = \frac{1 + 2 + 6 + x + y}{5}$$

Or

$$22 = 9 + x + y$$

Therefore

$$x + y = 13 \qquad (i)$$

Also

$$variance = 8.24 = \frac{1}{n}\sum_{i=1}^{5}(x_i - \bar{x})^2$$

i.e.,

$$8.24 = \frac{1}{5}[(3.4)^2 + (2.4)^2 + (1.6)^2 + x^2 + y^2 - 2 \times 4.4(x + y) + 2 \times (4.4)^2]$$

Or

$$41.2 = 11.56 + 5.76 + 2.56 + x^2 + y^2 - 8.8 \times 13 + 38.72$$

Therefore

$$x^2 + y^2 = 97 \qquad (ii)$$

But from (i), we have

$$x^2 + y^2 + 2xy = 169 \qquad (iii)$$

From (ii) and (iii), we have

$$2xy = 72 \qquad (iv)$$

Subtracting (iv) from (ii), we get

$$x^2 + y^2 - 2xy = 97 - 72 \quad i.e., \quad (x-y)^2 = 25$$

Or

$$x - y = \pm 5 \qquad (v)$$

So, from (i) and (v), we get

$$x = 9, \qquad y = 4 \ when \ x - y = 5$$

Or

$$x = 4, \qquad y = 9 \ when \ x - y = -5$$

Thus, the remaining observations are 4 and 9.

Example 18: If each of the observations $x_1, x_2, x_3, \cdots, x_n$ is increased by 'a', where a is a negative or positive number, show that the variance remains unchanged.

Solution: Let $\bar{x}$ be the mean of $x_1, x_2, x_3, \cdots, x_n$. Then the variance is given by

$$\sigma_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

If 'a' is added to each observation, the new observations will be

$$y_i = x_i + a \qquad (i)$$

Let the mean of the new observations be $\bar{y}$. Then

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_i + a) = \frac{1}{n}\left[\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} a\right] = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{na}{n} = \bar{x} + a$$

i.e.

$$\bar{y} = \bar{x} + a \qquad (ii)$$

Thus, the variance of the new observations

$$\sigma_2^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i + a - \bar{x} - a)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sigma_1^2$$

$$[using \ (i) \ and \ (ii)]$$

Thus, the variance of the new observations is same as that of the original observations.

Note: We may note that adding (or subtracting) a positive number to (or from) each observation of a group does not affect the variance.

Example 19: The mean and standard deviation of 100 observations were calculated as 40 and 5.1, respectively by a student who took by mistake 50 instead of 40 for one observation. What are the correct mean and standard deviation?

Solution: Given that number of observations $(n) = 100$.

Incorrect mean $(\bar{x}) = 40$.

Incorrect standard deviation $(\sigma) = 5.1$.

We know that

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

i.e.,

$$40 = \frac{1}{100}\sum_{i=1}^{100} x_i \quad or \quad \sum_{i=1}^{100} x_i = 4000$$

i.e., incorrect sum $= 4000$.

Thus the correct sum of observations=incorrect sum-50=4000-50+40=3990.

Hence correct mean $= \frac{correct\ sum}{100} = \frac{3990}{100} = 39.9$

Also

$$standard\ deviation\ \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2}$$

i.e.,

$$5.1 = \sqrt{\frac{1}{100} \times incorrect \sum_{i=1}^{n} x_i^2 - (40)^2}$$

Or

$$26.01 = \frac{1}{100} \times incorrect \sum_{i=1}^{n} x_i^2 - 1600$$

Therefore,

$$incorrect \sum_{i=1}^{n} x_i^2 = 100(26.01 + 1600) = 162601$$

Now

$$correct \sum_{i=1}^{n} x_i^2 = incorrect \sum_{i=1}^{n} x_i^2 - (50)^2 + (40)^2 = 162601 - 2500 + 1600$$
$$= 161701$$

Therefore, correct standard deviation

$$= \sqrt{\frac{correct\ \sum_{i=1}^{n} x_i^2}{n} - (correct\ mean)^2} = \sqrt{\frac{161701}{100} - (39.9)^2}$$
$$= \sqrt{1617.01 - 1592.01} = \sqrt{25} = 5.$$

Miscellaneous Exercise on Chapter 7

1. The mean and variance of eight observations are 9 and 9.25, respectively. If six of the observations are 6, 7, 10, 12, 12 and 13, find the remaining two observations.
2. The mean and variance of 7 observations are 8 and 16, respectively. If five of the observations are 2, 4, 10, 12, 14. Find the remaining two observations.

3. The mean and standard deviation of six observations are 8 and 4, respectively. If each observation is multiplied by 3, find the new mean and new standard deviation of the resulting observations.
4. Given that $\bar{x}$ is the mean and σ^2 is the variance of n observations $x_1, x_2, x_3, \cdots, x_n$. Prove that the mean and variance of the observations $ax_1, ax_2, ax_3, \cdots, ax_n$ are ax and $a^2\sigma^2$, respectively, $(a \neq 0)$.
5. The mean and standard deviation of 20 observations are found to be 10 and 2, respectively. On rechecking, it was found that an observation 8 was incorrect. Calculate the correct mean and standard deviation in each of the following cases:
    - If wrong item is omitted.
    - If it is replaced by 12.
6. The mean and standard deviation of marks obtained by 50 students of a class in three subject, mathematics, physics, and chemistry are given below:

Subjects	Mathematics	Physics	Chemistry
Mean	42	32	40.9
Standard deviation	12	15	20

Which of the three subjects shows the highest variability in marks and which shows the lowest?

7. The mean and standard deviation of a group of 100 observations were found to be 20 and 3, respectively. Later on it was found that three observations were incorrect, which were recorded as 21, 21 and 18. Find the mean and standard deviation if the incorrect observations are omitted.

Chapter Summary

Measures of dispersion: Range, Quartile deviation, mean deviation, variance, standard deviation are measures of dispersion.

$$Range = Maximum\ Value - Minimum\ Value$$

Mean deviation for ungrouped data:

$$M.D.(\bar{x}) = \frac{\sum|x_i - \bar{x}|}{n}, \qquad M.D.(M) = \frac{\sum|x_i - M|}{n}$$

Mean deviation for grouped data:

$$M.D.(\bar{x}) = \frac{\sum f_i|x_i - \bar{x}|}{N}, \qquad M.D.(M) = \frac{\sum f_i|x_i - M|}{N}, \qquad where\ N = \sum f_i$$

Variance and standard deviation for ungrouped data:

$$\sigma^2 = \frac{1}{n}\sum(x_i - \bar{x})^2, \qquad \sigma = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$$

Variance and standard deviation of a discrete frequency distribution:

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

Variance and standard deviation of a continuous frequency distribution:

$$\sigma^2 = \frac{1}{N}\sum f_i(x_i - \bar{x})^2, \qquad \sigma = \frac{1}{N}\sqrt{N\sum f_i x_i^2 - \left(\sum f_i x_i\right)^2}$$

Shortcut method to find variance and standard deviation:

$$\sigma^2 = \frac{h^2}{N^2}\left[N\sum f_i y_i^2 - \left(\sum f_i y_i\right)^2\right], \qquad \sigma = \frac{h}{N}\sqrt{N\sum f_i y_i^2 - \left(\sum f_i y_i\right)^2}$$

$$where\ y_i = \frac{x_i - A}{h}$$

Coefficient of variation (C.V.) $= \frac{\sigma}{x} \times 100,\ \bar{x} \neq 0.$

For series with equal means, the series with lesser standard deviation is more consistent or less scattered.

Historical Note: 'Statistics' is derived from the Latin word 'status' which mean a political state. This suggests that statistics is as old as human civilization. In the year 3050 B.C., perhaps the first census was held in Egypt. In the Middle East and south Asia also, about 2000 years ago, they had an efficient system of collecting administrative statistics, particularly, during the regime of Chandra Gupta Maurya (324-300 B.C.) in India. The system of collecting data related to births and deaths is mentioned in Kautilya's Arthshastra (around 300 B.C.). A detailed account of administrative survey conducted during Akbar Shah's regime is given in Ain-I-Akbari written by Abul Fazl.

References

Francis, A. (2004). *Business Mathematics and Statistics*. Birmingham: Thomson Learning.

Keller, G. (2014). *Statistics for Management and Economics*. Canada: Cengage Learning.

Kokoska, S. (2015). *Introductory Statistics*. New York: W. H. Freeman and Company.

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2010). *Statistical Techniques in Business and Economics*. New York: McGraw-Hill/Irwin.

Weiss, N. A. (2012). *Introductory Statistics*. Boston: Addison-Wesley.

Wiley, J. (2014). *1,001 Statistics Practice Problems For Dummies*. New Jerey: John Wiley & Sons, Inc.