

DATA HANDLING

INTRODUCTION OF DATA HANDLING

1. The word data means information (its exact dictionary meaning is: given facts).
Statistical data are of two types
(i) Primary data (ii) Secondary data
2. When an investigator collects data himself with a definite plan or design in his (her) mind, it is called Primary data.
3. Data which are not originally collected rather obtained from published or unpublished sources are known as Secondary data.
4. After collection of data, the investigator has to find ways to condense them in tabular form in order to study their salient features. Such an arrangement is called Presentation of data.
5. Raw data when put in ascending or descending order of magnitude is called an array or arranged data.
6. The number of times an observation occurs in the given data is called frequency of the observation.
7. Classes/class intervals are the groups in which all the observations are divided.
8. Suppose class-interval is 10-20, then 10 is called lower limit and 20 is called upper limit of the class
9. Mid-value of class-interval is called Class-mark
$$\text{Class-mark} = \frac{\text{Lower limit} + \text{upper limit}}{2}$$
$$\text{Class-mark} = \text{lower limit} + \frac{1}{2} (\text{difference between the upper and lower limits})$$
10. If the frequency of first class interval is added to the frequency of second class and this sum is added to third class and so on then frequencies so obtained are known as Cumulative Frequency (c.f.).
11. There are two types of cumulative frequencies (a) less than, (b) greater than

IMPORTANT FACTS AND FORMULAE

1. **Experiment :** An operation in which can produce some well-defined outcomes is called an experiment.
2. **Random Experiment :** An Experiment in which all possible outcomes are known and the exact output cannot be predicated in advance, is called a random experiment.
3. **Examples of Performing a Random Experiment :**
 - (i) Rolling an unbiased dice.
 - (ii) Tossing a fair coin.
 - (iii) Drawing a card from a pack of well-shuffled cards.
 - (iv) Picking up a ball of certain colour from a bag containing balls of different colours.
4. **Details :**
 - (i) When we throw a coin. Then either a Head (H) or a Tail (T) appears.
 - (ii) A dice is a solid cube, having 6 faces, marked 1, 2, 3, 4, 5, 6 respectively. When we throw a die, the outcome is the number that appears on its upper face.
 - (iii) A pack of cards has 52 cards. It has 13 cards of each suit, namely Spades, Clubs, Hearts and Diamonds. Cards of spades and clubs are black cards. Cards of hearts and diamonds are red cards.
There are 4 honours of each suit.
These are Aces, Kings, Queens and Jacks.
These are called face cards.
5. **Sample Space :** When we perform an experiment, then the set S of all possible outcomes is called the Sample Space.
6. **Examples of Sample Spaces :**
 - (i) In tossing a coin, $S = \{H, T\}$.
 - (ii) If two coins are tossed, then $S = \{HH, HT, TH, TT\}$.
 - (iii) In rolling a dice, we have, $S = \{1, 2, 3, 4, 5, 6\}$.
7. **Event :** Any subset of a sample space is called an event.

8. Probability of Occurrence of an Event

Let S be the sample space and let E be an event.

$$\text{Then, } E \subseteq S \quad \therefore P(E) = \frac{n(E)}{n(S)}.$$

DATA

The collection of facts which are expressed numerically with the specific purpose is called data.

Each numerical fact of this type is known as an observation.

COLLECTION OF DATA

On the basis of methods of collection data can be divided in to two categories :

(a) Primary data (b) Secondary Data

(a) **Primary data** : The data which are collected by the investigator itself are called primary data.

(b) **Secondary Data** : Data already collected earlier and other investigator use it for his investigation then it is called secondary data for investigator. It may be published or unpublished.

Raw Data : A collection of observations gathered initially is called raw data.

Arrayed Data : The data arranged in an order descending or ascending order is called arrayed data.

Range : The difference between the highest and the lowest value of the observations in a data is called the range of the data.

e.g., Let us suppose 25 students obtained the marks

Highest marks obtained = 88

Lowest marks obtained = 30

Range = $88 - 30 = 58$

Ex.1 Given below are the marks (out of 100) in mathematics obtained by 20 students of a class in an annual examination.

23 75 56 42 70 84 92 51 40 63
87 58 35 80 14 63 49 72 66 61

Arrange the above data in ascending order and find

- (i) the lowest marks obtained.
- (ii) the highest marks obtained.
- (iii) the range of the given data.

Sol. Arranging the above data in ascending order, we get :

14 23 35 40 42 49 51 56 58 61
63 63 66 70 72 75 80 84 87 92

- (i) Lowest marks obtained = 14.
- (ii) Highest marks obtained = 92
- (iii) Range of the given data = $(92 - 14) = 78$

ARITHMETIC MEAN

Mean of n observation $x_1, x_2, x_3, \dots, x_n$ is given by $\bar{X} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$

$$\text{Mean} = \frac{\text{Sum of all observations}}{\text{number of observation}}$$

FREQUENCY DISTRIBUTION

Frequency : In a given data the number of times a particular observation occurs is called its frequency. It is usually denoted by (**f_i**).

Frequency Distribution tables : A table showing the frequencies of the various observation of a data is called a frequency distribution table of simply a frequency table.

Tally marks :

- (i) When the number of observations is large, we make use of tally marks (||||) to find the frequencies.
- (ii) Tallies are usually marked in a bunches of five. [The fifth tally in a bunch is usually marked diagonally across the earlier (||||) The fifth one crossing the other four diagonally (||||) which represents five.

Ex.2 The marks scored by 30 students of IX class, of a school in the first test of Mathematics out of 50 marks are as follows :

6 32 10 17 22 28 0 48 6 22
 32 6 36 25 48 10 32 48 28 22
 22 22 28 26 17 36 10 22 28 0

Sol. From the observation of given raw data, it is difficult to judge the standard of the class correctly but if we prepare a table showing how many students scored 0, how many 6, how many 12, etc. then it becomes more easy to understand the standard of the class. The number of times a mark is repeated is called its frequency. It is denoted by f .

The following table is obtained from the above data :

Marks Obtained	Tally mark	Frequency
0	II	2
6	III	3
10	III	3
17	II	2
22	IIII	6
25	I	1
26	I	1
28	IIII	4
32	III	3
36	II	2
48	III	3

Ex.3 In a study of number of accidents per day, the observations for 30 days were obtained as follows :

4, 3, 5, 6, 4, 3, 2, 5, 4, 2, 6, 2, 1, 2, 2, 0, 5, 4, 6, 1, 3, 0, 5, 3, 6, 1, 5, 5, 2, 6.

Prepare a frequency table.

Sol. Arranging the data in ascending order, we get

0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6

Now, we represent the above data in the tubular form as shown below :

No. of accidents	Tally mark	Frequency
0	II	2
1	III	3
2		6
3		4
4		4
5		6
6		5
	Total	30

NOTE : Such type of frequency distribution is called ungrouped frequency distribution.

GROUPING OF DATA

When the list of observation is long, the data is usually organised in to groups called class intervals and the data so obtained is called a grouped data.

Presentation of data in groups along with the frequency of each group is known as the frequency distribution of the grouped data.

SOME IMPORTANT DEFINITION

Lower limit & Upper limit : The lowest value of class-interval is called its lower limit and the upper value of a class-interval is called its upper limit.

Class size : The difference between the upper and lower class limits is called the class size.

e.g., The size of class intervals = upper limit – lower limit

$$5 - 0 = 5, 30 - 20 = 10, 40 - 30 = 10$$

Class mark : The mid value of a class-interval is called its class mark.

$$\text{Class mark} = \frac{\text{upper limit} + \text{lower limit}}{2}$$

Thus, the class marks of 0–5 is = 2.5, The class mark of 5–10 is $\frac{5+10}{2} = 7.5$ etc.

Class interval : In the above frequency distribution.

$$\text{Class interval} = \frac{\text{Range}}{\text{Number of classes}} \quad (\text{it is generally denoted by } h.)$$

If x be the mid value and h be the class interval, then the class limits are

$$\left(x - \frac{h}{2}, x + \frac{h}{2}\right)$$

Ex.4 The class marks of a distribution are : 5, 15, 25, 35, 45. Find the classes and the class size.

Sol. Here the class size = $15 - 5 = 10$

Lower limit of first class = $5 - \frac{10}{2} = 0$, Upper limit of first class = $5 + \frac{10}{2} = 10$

The classes are 0–10, 10–20, 20–30, 30–40, 40–50 and class size is 10.

GRAPHICAL METHOD OF REPRESENTING DATA


Some of the forms to represent data graphically are















(a) A pictograph (b) A bar graph (c) Double bar graph (d) Histogram

(a) A Pictograph

In Pictograph, we represent data with the help of symbol.



 = 50 Cycle \rightarrow one symbol stands for 50 Cycle.

Amit	   	50×4	$= 200$
Sachin	  	50×3	$= 150$
Manish	    	50×5	$= 250$
Ankit	 	$50 \times ?$	$= ?$

- (i) How many Cycle does Ankit has?
- (ii) Who has maximum Cycle?

(iii) Who has minimum Cycle?

(b) A Bar Graph

A bar graph is a pictorial representation of numerical data in the form of rectangles (or bars) of equal width and varying heights.

These rectangles are drawn either vertically or horizontally, keeping equal space between them. The height (or length) of a rectangle depends upon the numerical value it represents.

Note : Bar graphs of grouped data are also called histograms.

ALGORITHM TO DRAW A BAR GRAPH

We can draw the graph by following the steps given below.

Step-I : On a graph paper, draw a horizontal line OX and a vertical line OY. These lines are called the x-axis and the y-axis respectively.

Step-II : Mark points at equal intervals along the x-axis. Below these points write the names of the data items whose values are to be plotted.

Step-III : Choose a suitable scale. On that scale determine the heights of the bars for the given numerical values.

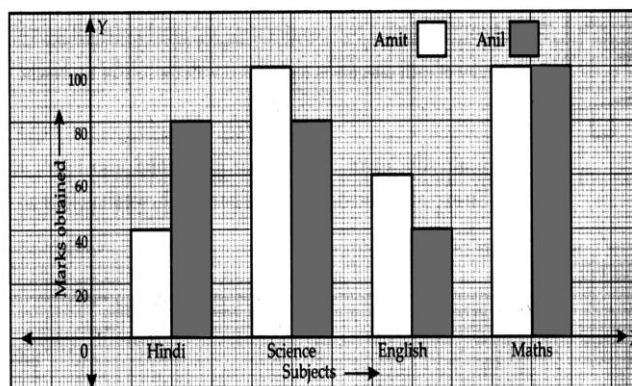
Step-IV : Mark off these heights parallel to the y-axis from the points taken in Step-II.

Step-V : On the x-axis, draw bars of equal width for the heights marked in Step-IV. The bars should be centrad on the points marked on the x-axis. These bars represent the given numerical data.

(c) Double Bar Graph

Double bar graph is used to compare the two sets of data simultaneously.

Ex.5



- (i) What is the information given by the double bar graph?
- (ii) In which subject both students have scored equal marks?
- (iii) In Science, whose performance is better?

Sol. (i) The double bar graph represents the marks obtained by two students Amit and Anil.

(ii) In Maths, both students have scored equal marks.

(iii) In Science, Amit's performance is better.

(d) Histogram

A histogram or frequency histogram is a graphical representation of a frequency distribution in the form of rectangles with class intervals as bases and the corresponding frequencies as the heights such that there is no gap between any two successive rectangles.

In drawing the histogram of a continuous grouped frequency distribution, we use the following algorithm.

ALGORITHM

- Step-I :** Take a graph paper and draw two perpendicular lines, one horizontal and one vertical, intersecting at O (say). Mark them as OX and OY.
- Step-II :** Take horizontal line OX as X-axis and vertical line OY as Y-axis.
- Step-III :** Choose a suitable scale for X-axis and along X-axis represent class-limits.
- Step-IV :** Choose a suitable scale for Y-axis and mark frequencies along Y-axis.
- Step-V :** Construct rectangles with respective class intervals as the bases and the corresponding class frequencies as heights.

We give a break called kink indicated by and then start drawing histogram.

NOTE: It should be noted that the scale for X-axis may not be same as the scale for Y-axis. The selection of scale depends upon our convenience and the type of data.

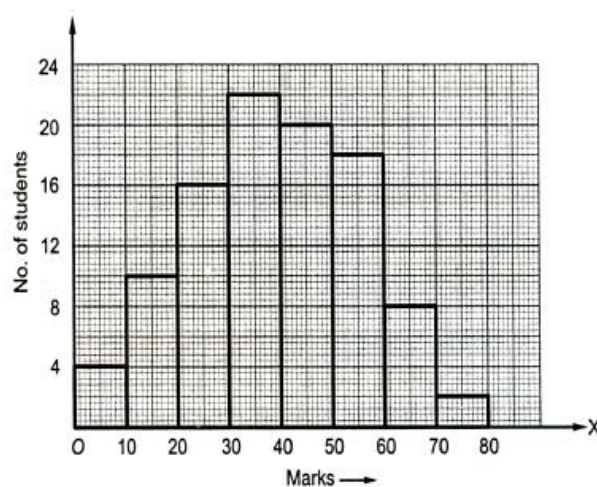
Ex.6 The following table gives the marks scored by 100 students in an entrance examination.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Number of students (Frequency)	4	10	16	22	20	18	8	2

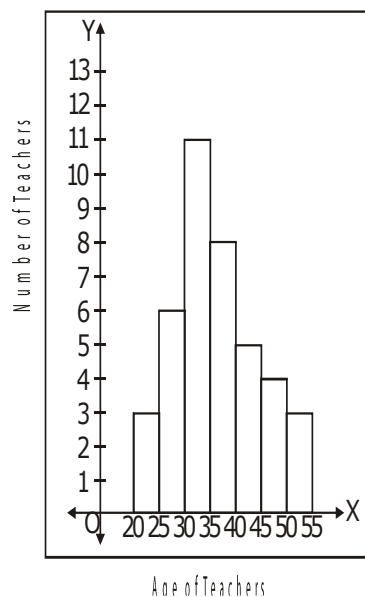
Represent this data in the form of a histogram.

Sol. We represent the class limits along X-axis on a suitable scale and the frequencies along Y-axis on a suitable scale.

Taking class-intervals as bases and the corresponding frequencies as heights, we construct rectangles to obtain the histogram of the given frequency distribution as shown in Fig.



Ex.7 Read the following histogram and answer the questions given at the end.



- (i) What information is depicted by the histogram?
- (ii) What is the most common age group? How many teachers are there in this group?
- (iii) What is the number of teachers who are more than 35 years of age but less than 40 years?
- (iv) How many teachers are 40 years or older?
- (v) Which groups contain the same number of teachers? (vi) What are the class marks of the classes?

- Sol.**
- (i) The histogram shows the distribution of ages (in year) of 40 teachers in a school.
 - (ii) The most common age group is 30–35 and the number of teachers in this age group is 11.
 - (iii) The number of teachers who are more than 35 years of age but less than 40 years is 8.
 - (iv) There are $12(5 + 4 + 3)$ teachers who are 40 years or older.
 - (v) The group 20–25 and 50–55 contain the same number of teachers.
 - (vi) The class mark of 20–25 is $\frac{20+25}{2}$, i.e. 22.5

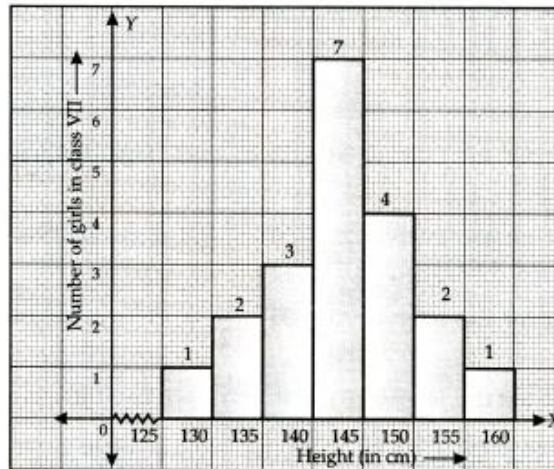
Similarly, the class mark of 25–30 is 27.5 the class mark of 30–35 is 32.5

the class mark of 35–40 is 37.5 the class mark of 40–45 is 42.5

the class mark of 45–50 is 47.5 the class mark of 50–55 is 52.5

Ex.8 Given below is the histogram, observe it and answer the following questions :

- (i) What information is being given by histogram?
- (ii) Which group contains maximum girls?
- (iii) How many girls have a height of 145 cm and above?



- Sol.**
- (i) Height (in cm) of 20 girls of class VII.
 - (ii) Group (140-145) contains maximum girls.
 - (iii) The number of girls having a height of 145 cm and more = $4 + 2 + 1 = 7$.

DIFFERENCES BETWEEN A BAR GRAPH AND A HISTOGRAM

- (i) In a bar graph the bars are at a distance from one another, while in a histogram the bars (rectangles) touch one another.
- (ii) In a bar graph we generally have a scale for either the X-axis or Y-axis (generally Y axis as vertical bars are more popular), but in a histogram, we have scale for both X axis and Y-axis and both scales need not be the same.
- (iii) In a bar graph, one axis may not have numerical values but have names, subjects, flavours etc. along it, but in histogram we display numerical values along both the axis.