STATISTICS

INTRODUCTION

Statistics is a broad mathematical discipline which studies ways to collect, summarize and draw conclusions from data It is applicable to a wide variety of academic fields from the physical & social sciences to the humanities as well as to business, government & industry.

In short, the study of data, how to collect , summarize & present it, is called statistics.

USEFUL TERMS

1. Limit of the Class

The starting and end values of each class are called **Lower** and **Upper limits**.

2. Class Interval

The difference between upper and lower boundary of a class is called class interval or size of the class.

3. Primary and Secondary Data

The data collected by the investigator himself is known as the **primary data**, while the data collected by a person other than the investigator is known as the **secondary data**.

4. Variable or Variate

A characteristics that varies in magnitude from observation to observation. e.g. weight, height, income, age, etc are variables.

5. Frequency

The number of times an observation occurs in the given data, is called the frequency of the observation.

6. Discrete Frequency Distribution

A frequency distribution is called a distance frequency distribution, if data are presented in such a way that exact measurements of the units are clearly shown.

7. Continuous Frequency Distribution

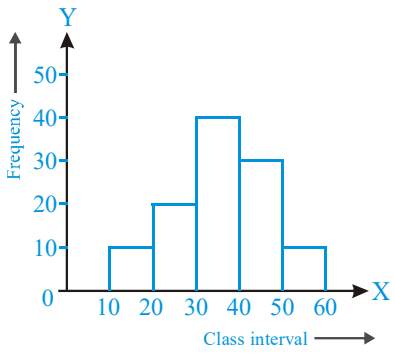
A frequency distribution which data are arranged in classes (or groups) which are not exactly measurable.

8. Cumulative Frequency Distribution

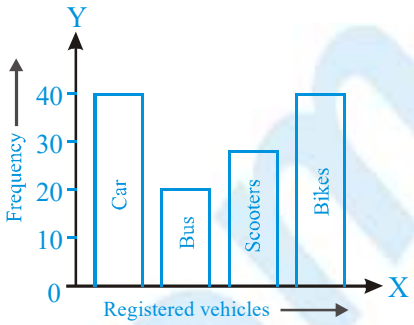
Suppose the frequencies are grouped frequencies or class frequencies. If however, the frequency of the first class is added to that of the second and this sum is added to that of the third and so on, then the frequencies, so obtained are known as cumulative frequencies (cf).

(i) **Histogram:** to draw the histogram of a given continuous frequency distribution, we first mark off all the class intervals along X-axis on a suitable scale. On each of these class intervals on the horizontal axis, we eract (vertical) a rectangle whose height is proportional to the frequency of that particular class, so that the area of the rectangle is proportional to the frequency of the class.

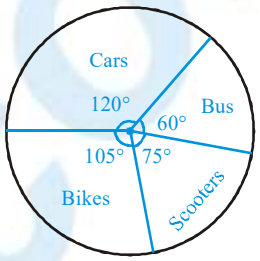
If however the classes are of unequal width, then the height of the rectangles will be proportional to the ratio of the frequencies to the width of the classes.



(ii) **Bar Diagrams:** In bar diagrams, only the length of the bars are taken into consideration. To draw a bar diagram, we first mark equal lengths for the different classes on the horizontal axis, i.e,. X-axis. On each of these lengths on the horizontal axis, we erect (vertical) a rectangle whose heights are proportional to the frequency of the class.



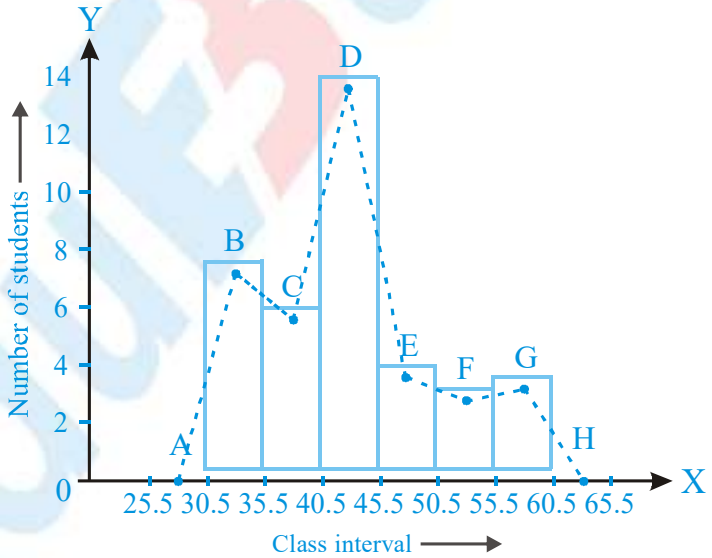
(iii) **Pie Diagrams:** Pie are used to represent a relative frequency distribution. A pie diagram consists of a circle divided into as many sectors as there are classes in a frequency distribution. The area of each sector is proportional to the relative frequency of the class.



Now, we make angles at the centre proportional to the relative frequencies. And in order to get the angles of the desired sectors, we divide 360° in the proportion of the various relative frequencies, i.e,.

$$\text{Central angle} = \left[\frac{\text{Frequence} \times 360°}{\text{Total frequency}}\right]$$

(iv) **Frequency Polygon** To draw the frequency polygon of an ungrouped frequency distribution, we plot the points with abscissas as the variate values and the ordinate as the corresponding frequencies. These plotted points are joined by straight lines to obtain the frequency polygon.

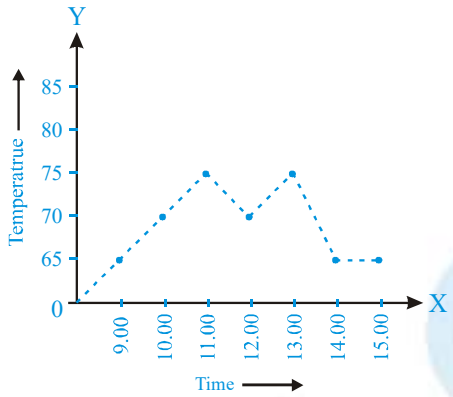


(v) Cumulative Frequency Curve (Ogive): The term ogive is pronounced as ogive. It is a shape consisting of a concave arc flowing into a convex arcs. i.e., forming as S-shaped curve with vertical ends.

There are two methods of constructing an ogiven,

(i) 'less than' type ogive

(ii) 'more than' type ogive.



MEASURES OF CENTRAL TENDENCY

An average value or a central value of a distribution is the value of variable which is representative of the entire distribution, this representative value are called the measures of central tendency.

It can be divided into two groups :

(A) MATHEMATICAL AVERAGE

- **I.** Arithmetic mean or mean
- **II.** Geometric mean
- **III.** Harmonic mean

(B) POSITIONAL AVERAGE

- **I.** Median
- **II.** Mode

Arithmetic Mean

The AM, also called the average value is the quantity obtained by summing two or more numbers or variables and then dividing by the number of numbers or variables.

(i) AM of Ungrouped Distribution

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

$$\bar{x} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\Rightarrow \Sigma x_i = n\bar{x}$$

(ii) AM of Discrete Grouped Distribution

Let $x_1, x_2, \ldots, x_n$ be n observation and let $f_1, f_2, \ldots\ldots, f_n$ be their corresponding frequencies, then their mean

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n} \quad \text{or} \quad \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

(iii) AM of Continuous Grouped Distribution

Take mid points of given classes as x_i and use formula as given for discrete grouped data.

Short Cut Method

(a) Assumed Mean Method

If the values of x or (and) f are large, the calculation of arithmetic mean by the previous formula used, is quite tedious and time consuming. In such case we take the deviation from assumed mean A which is in the middle or just close to it in the data.

Then $\bar{x} = A + \frac{\sum f_i d_i}{\sum f_i}$ where $d_i = x_i - A$ = deviation for each observation

This method is nothing but shifting of origin from zero to the assumed mean on the number line.

(b) Step Deviation Method

Sometimes during the application of assumed mean method of finding the mean, the deviations d_i are divisible by a common number h(say). In such case the calculations are reduced to a great extent by using $u_i = \frac{x_i - A}{h}$, i = 1, 2, ..., n

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

This process is called change of scale on the number line.

(c) Weighted Mean

If $w_1, w_2, \ldots\ldots w_n$ are the weights assigned to the values $x_1, x_2, \ldots\ldots x_n$ respectively then their weighted mean is defined as

$$\text{Weighted mean} = \frac{w_1x_1 + w_2x_2 + \ldots\ldots + w_nx_n}{w_1 + \ldots\ldots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

(d) Combined Mean

If $\bar{x}_1$ and $\bar{x}_2$ be the means of two groups having n_1 and n_2 terms respectively then the mean (combined mean) of their composite group is given by

$$\text{combined mean} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

If there are more than two groups then, combined mean $= \frac{n_1\bar{x}_1 + n_1\bar{x}_2 + n_3\bar{x}_3 + \ldots.}{n_1 + n_2 + n_3 + \ldots.}$

PROPERTIES OF ARITHMETIC MEAN

(i) Sum of deviations of variate from their A.M. is always zero i.e. $\Sigma(x_i - \bar{x}) = 0$, $\Sigma f_i(x_i - \bar{x}) = 0$

(ii) Sum of square of deviations of variate from their A.M. is minimum i.e. $\Sigma(x_i - \bar{x})^2$ is minimum

(iii) If $\bar{x}$ is the mean of variate x_i then A.M. of $(x_i + \lambda) = \bar{x} + \lambda$

A.M. of $(\lambda x_i) = \lambda\bar{x}$

A.M. of $(ax_i + b) = a\bar{x} + b$ (where λ, a, b are constant)

(iv) A.M. is independent of change of assumed mean i.e. it is not effected by any change in assumed mean.

(v) If $\bar{x}$ is the mean of $x_1, x_2, \ldots\ldots, x_n$, then mean of $ax_1 + b$, $ax_2 + b$, ... $ax_n + b$ is $a\bar{x} + b$.

(vi) Arithmetic mean is dependent on change of origin & scale.

Ex. Find mean of data 2, 4, 5, 6, 8, 17.

Sol. Mean $\frac{2+4+5+6+8+17}{6} = 7$

Ex. Find the A.M. of the following freq. dist.

x_i	5	8	11	14	17
f_i	4	5	6	10	20

Sol. Here $N = \Sigma f_i = 4 + 5 + 6 + 10 + 20 = 45$

$\Sigma fx = (5 \times 4) + (8 \times 5) + (11 \times 6) + (14 \times 10) + (17 \times 20) = 606$

$\therefore \quad \bar{x} = \frac{\Sigma fx}{N} = \frac{606}{45} = 13.47$

Ex. Find the mean of the following freq. dist.

x_i	5	15	25	35	45	55
f_i	12	18	27	20	17	6

Sol. Let assumed mean $a = 35$, $h = 10$

here $N = \Sigma f_i = 100$, $u_i = \frac{(x_i - 35)}{10}$

$\therefore \quad \Sigma f_i u_i = (12 \times -3) + (18 \times -2) + (27 \times -1) + (20 \times 0) + (17 \times 1) + (6 \times 2) = -70$

$\therefore \quad \bar{x} = a + \left(\frac{\Sigma fu}{N}\right) h = 35 + \frac{(-70)}{100} \times 10 = 28$

Ex. Find the mean marks of students from the following cumulative frequency distribution :

Marks	Number of students	Marks	Number of students
0 and above	80	60 and above	28
10 and above	77	70 and above	16
20 and above	72	80 and above	10
30 and above	65	90 and above	8
40 and above	55	100 and above	0
50 and above	43		

Sol. Here we have, the cumulative frequency distribution. So, first we convert it into an ordinary frequency distribution. We observe that there are 80 students getting marks greater than or equal to 0 and 77 students have secured 10 and more marks. Therefore, the number of students getting marks between 0 and 10 is $80 - 77 = 3$.

Similarly, the number of students getting marks between 10 and 20 is $77 - 72 = 5$ and so on.

Thus, we obtain the following frequency distribution.

Marks	Number of students	Marks	Number of students
0 – 10	3	50 – 60	12
10 – 20	5	60 – 70	12
20 – 30	7	70 – 80	6
30 – 40	10	80 – 90	2
40 – 50	12	90 – 100	8

Now, we compute arithmetic mean by taking 55 as the assumed mean.

Computation of Mean

Marks	Mid-value (x_i)	Frequency (f_i)	$u_i = \frac{x_i - 55}{10}$	$f_i u_i$
0 – 10	5	3	–5	–15
10 – 20	15	5	–4	–20
20 – 30	25	7	–3	–21
30 – 40	35	10	–2	–20
40 – 50	45	12	–1	–12
50 – 60	55	15	0	0
60 – 70	65	12	1	12
70 – 80	75	6	2	12
80 – 90	85	2	3	6
90 – 100	95	8	4	32
Total		$\sum f_i = 80$		$\sum f_i u_i = -26$

We have,

$$N = \sum f_i = 80;\ \sum f_i u_i = -26,\ A = 55 \text{ and } h = 10$$

$$\therefore \quad \bar{X} = A + h\left(\frac{1}{N}\sum f_i u_i\right)$$

$$\Rightarrow \quad \bar{X} = 55 + 10 \times \frac{-26}{80} = 55 - 3.25 = 51.75 \text{ Marks}$$

Ex. Find the weighted mean of first n natural numbers when their weights are equal to their squares respectively

Sol. Weighted Mean $= \dfrac{1.1^2 + 2.2^2 + \ldots + n.n^2}{1^2 + 2^2 + \ldots + n^2} = \dfrac{1^3 + 2^3 + \ldots + n^3}{1^2 + 2^2 + \ldots + n^2} = \dfrac{[n(n+1)/2]^2}{[n(n+1)(2n+1)/6]} = \dfrac{3n(n+1)}{2(2n+1)}$

Ex. The mean income of a group of persons is Rs. 400 and another group of persons is Rs. 480. If the mean income of all the persons of these two groups is Rs. 430 then find the ratio of the number of persons in the groups.

Sol. Here $\bar{x}_1 = 400$, $\bar{x}_2 = 480$, $\bar{x} = 430$

$$\therefore \quad \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \quad \Rightarrow \quad 430 = \frac{400n_1 + 480n_2}{n_1 + n_2} \quad \Rightarrow \quad \frac{n_1}{n_2} = \frac{5}{3}$$

MERITS & DEMERITS OF ARITHMETIC MEAN

Merits

(i) It is rigidly defined.

(ii) It is based on all the observation taken.

(iii) It is calculated with reasonable ease.

(iv) It is least affected by fluctuations in sampling.

(v) It is based on each observation and so it is a better representative of the data.

(vi) It is relatively reliable

(vii) Mathematical analysis of mean is possible.

Demerits

(i) It is severely affected by the extreme values.

(ii) It cannot be represented in the actual data since the mean does not coincide with any of the observed value.

(iii) It cannot be computed unless all the items are known.

GEOMETRIC MEAN

(i) **GM of Ungrouped Distribution. :** If $x_1, x_2, ... x_n$ are n positive values of variate then their geometric mean G is given by

$$G = (x_1 \times x_2 \times ..... \times x_n)^{1/n}$$

$$\Rightarrow \quad G = \text{antilog}\left[\frac{1}{n}\sum_{i=1}^{n}\log x_i\right]$$

(ii) **GM of Frequency Distribution :** If $x_1, x_2, ... x_n$ are n positive values with corresponding frequencies $f_1, f_2, ... f_n$ resp. then their G.M.

$$G = (x_i^{f_1} \times x_2^{f_2} \times ..... \times x_n^{f_n})^{1/N}$$

$$\Rightarrow \quad G = \text{antilog}\left[\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right]$$

If G_1 and G_2 are geometric means of two series which containing n_1 and n_2 positive values resp. and G is geometric mean of their combined series then

$$G = (G_1^{n_1} \times G_2^{n_2})^{\frac{1}{n_1 + n_2}}$$

$$\Rightarrow \quad G = \text{antilog}\left[\frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}\right]$$

Ex. Find the G.M. of $1, 2, 2^2, \ldots, 2^n$

Sol. G.M. $= (1.2.2^2.....2^n)^{\frac{1}{n+1}}$

$$= \left[2^{\frac{n(n+1)}{2}}\right]^{\frac{1}{n+1}} = 2^{n/2}$$

HARMONIC MEAN

(i) HM of Ungrouped Distribution : If $x_1, x_2, \ldots x_n$ are n non-zero values of variate then their harmonic mean H is defined as

$$H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots.+\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

(ii) HM of Frequency Distribution : If $x_1, x_2, \ldots x_n$ are n non-zero values of variate with corresponding frequencies $f_1, f_2, \ldots f_n$ respectively the their H.M.

$$H = \frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\ldots.+\frac{f_n}{x_n}} = \frac{N}{\sum_{i=1}^{n}\frac{f_i}{x_i}}$$

MEDIAN

Median is the middle most or the central value of the variate in a set of observations when the observations are arranged either in ascending or in descending order of their magnitudes. It divides the arranged series in two equal parts.

(i) Median of Ungrouped Distribution

Let $x_1, x_2, \ldots\ldots, x_n$ be n observations arranged in ascending or descending order. Then

$$\text{Median (M)} = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ observation if n is odd}$$

$$\text{Median (M)} = \frac{\left(\frac{n}{2}\right)^{th}\text{observation} + \left(\frac{n}{2}+1\right)^{th}\text{observation}}{2} \text{ if n is even}$$

(ii) Median of Ungrouped Frequency Distribution

Find the cumulative frequency (C.F.)

$$\text{Median (M)} = \text{Value of } \left(\frac{N+1}{2}\right)^{th} \text{ observation if N is odd}$$

$$\text{Median (M)} = \frac{\left(\frac{N}{2}\right)^{th}\text{observation} + \left(\frac{N}{2}+1\right)^{th}\text{observation}}{2} \text{ if N is even}$$

where $N = \sum_{i=1}^{n} f_i$

(iii) Median of Grouped Frequency Distribution

Let the number of observation be N. Prepare the cumulative frequency table. Find the median class i.e. the class in which the observation whose cumulative frequency is equal to or just greater than $\frac{N}{2}$ lies.

The median value is given by the formula : Median (M) $= \bullet + \left| \frac{\left(\frac{N}{2}\right) - c}{f} \right| \times h$ where

N = total frequency = Σf_i

$\bullet$ = lower limit of median class

f = frequency of the median class

c = cumulative frequency of the class preceding the median class

h = class interval (width) of the median class

Ex. Find the median of observations 4, 6, 9, 4, 7, 8, 10

Sol Values in ascending order are 4, 4, 6, 7, 8, 9, 10

here n = 7 so $\frac{n+1}{2} = 4$

so median = 4th observation = 7

Ex. Obtain the median for the following frequency distribution :

x	1	2	3	4	5	6	7	8	9
f	8	10	11	16	20	25	15	9	6

Sol

x	f	cf
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
	N = 120	

Here, $N = 120 \rightarrow \frac{N}{2} = 60$

We find that the cumulative frequency just greater than $\frac{N}{2}$ i.e., 60 is 65 and the value of x corresponding to 65 is 5.

Therefore, Median = 5.

Ex. Find the median of following freq. dis.

class	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
f	8	30	40	12	10

Sol.

class	f	c.f
0 - 10	8	8
10 - 20	30	38
20 - 30	40	78
30 - 40	12	90
40 - 50	10	100

Here $\frac{N}{2} = \frac{100}{2} = 50$ which lies in the value 78 of c.f. hence corresponding class of this c.f. is 20-30 is the median class. So

- $l = 20, f = 40, F = 38, h = 10$

$\therefore$ Median $= l + \left(\frac{\frac{N}{2} - F}{f}\right) \times h = 20 + \frac{(50 - 38)}{40} \times 10 = 23$

MERITS AND DEMERITS OF MEDIAN

Merits

(i) It is easy to compute and understand.

(ii) It is well defined an ideal average should be

(iii) It can also be computed in case of frequency distribution with open ended classes.

(iv) It is not affected by extreme values.

(v) It can be determined graphically.

(vi) It is proper average for qualitative data where items are not measured but are scored.

Demerits

(i) For computing median data needs to be arranged in ascending or descending order.

(ii) It is not based on all the observations of the data.

(iii) It cannot be given further algebraic treatment.

(iv) It is affected by fluctuations of sampling.

(v) It is not accurate when the data is not large.

(vi) In some cases median is determined approximately as the mid-point of two observations whereas for mean this does not happen.

MODE

Mode is that value in a series which occurs most frequently. In a frequency distribution, mode is that variate which has the maximum frequency.

(i) Mode of Ungrouped Distribution

In the case of individual series, the value which is repeated maximum number of times is the mode of the series.

(ii) Mode of Ungrouped Frequency Distribution

In the case of discrete frequency distribution, mode is the value of the variate corresponding to the maximum frequency.

(iii) Mode of Grouped Frequency Distribution

First find the modal class i.e. the class which has maximum frequency. The modal class can be determined either by inspecting or with the help of grouping.

The mode is given by the formula :

$$\text{Mode} = \bullet + \frac{f_1 - f_{0}}{2f_1 - f_{1-1} - f_{1+1}} \times h$$

where $\bullet$ = lower limit of the modal class

h = width of the modal class

f_1 = frequency of the modal class

f_{1-1} = frequency of the class preceding modal class

f_{1+1} = frequency of the class succeeding modal class

Ex. Find mode of data 2, 4, 6, 8, 8, 12, 17, 6, 8, 9.

Sol. 8 occurs maximum number of times so mode = 8

Ex. Find the mode of the following frequency dist

Class	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
f_i	2	18	30	45	35	20	6	3

Sol. Here the class 30–40 has maximum freq. so this is the modal class

$\bullet = 30, f_1 = 45, f_{1-1} = 30, f_{1+1} = 35, h = 10$

$$\therefore \quad \text{Mode} = \bullet + \frac{f_1 - f_{1-1}}{2f_1 - f_{1-1} - f_{1+1}} \times h = 30 + \frac{45 - 30}{2 \times 45 - 30 - 35} \times 10 = 36$$

MERITS, DEMERITS OF MODE

Merits

(i) It is readily comprehensible and easy to compute. In some case it can be computed merely by inspection.

(ii) It is not affected by extreme values. It can be obtained even if the extreme values are not known.

(iii) Mode can be determined in distributions with open classes.

(iv) Mode can be located on graph also.

Demerits

(i) It is ill-defined. It is not always possible to find a clearly defined mode. In some cases, we may come across distributions with two modes. Such distributions are called bimodal. If a distribution has more than two modes, it is said to be multimodal.

(ii) It is not based upon all the observation.

(iii) Mode can be calculated by various formulae as such the value may differ from one to other. Therefore, it is not rigidly defined.

(iv) It is affected to a greater extent by fluctuations of sampling.

USES OF MODE

Mode is used by the manufacturers of ready-made garments, shoes and accessories in common use etc. The readymade garment manufacturers made those sizes more which are used by most of the persons than other sizes. Similarly, the makers of shoes will make that size maximum which the majority people use and others in less quantity.

Relationship Between Mean, Mode and Median

(i) In symmetrical distribution, Mean = Mode = Median

(ii) In skew (moderately asymmetrical) distribution, Mode = 3 Median – 2 Mean

GRAPHICAL REPRESENTATION OF DATA

Usually comparisons among the individual items are best shown by means of graphs. The representation then becomes easier to understand than the actual data. We shall study the following graphical representations in this section

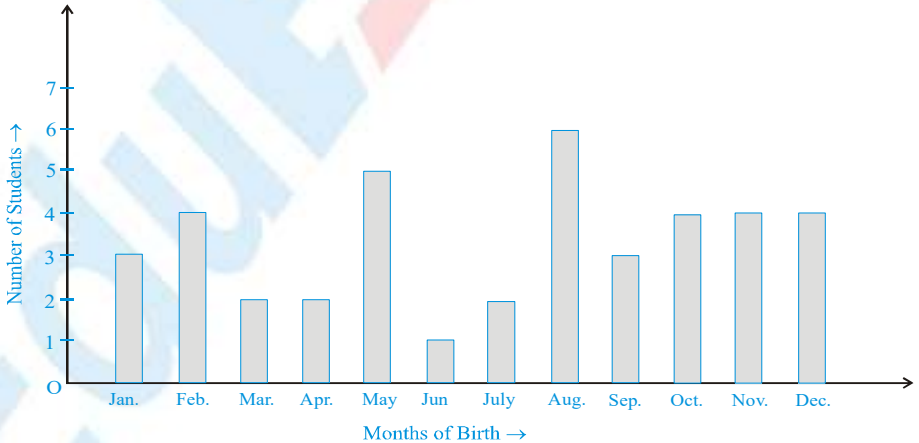
(A) Bar graphs

(B) Histograms of uniform width and of varying widths

(C) Frequency polygons

(A) Bar Graphs

A bar graph is a pictorial representation of data in which usually bars of uniform width are drawn with equal spacing between them on one axis (say, x-axis), depicting the variable. The values of the variable are shown on the other axis (say, the y-axis) and the heights of the bars depend on the values of the variable. For example, in a particular section of class IX, 40 students were asked about the months of their birth and the following graph was prepared for the data so obtained :



The variable here is the 'month of birth' and the value of the variable is the 'Number of students born'.

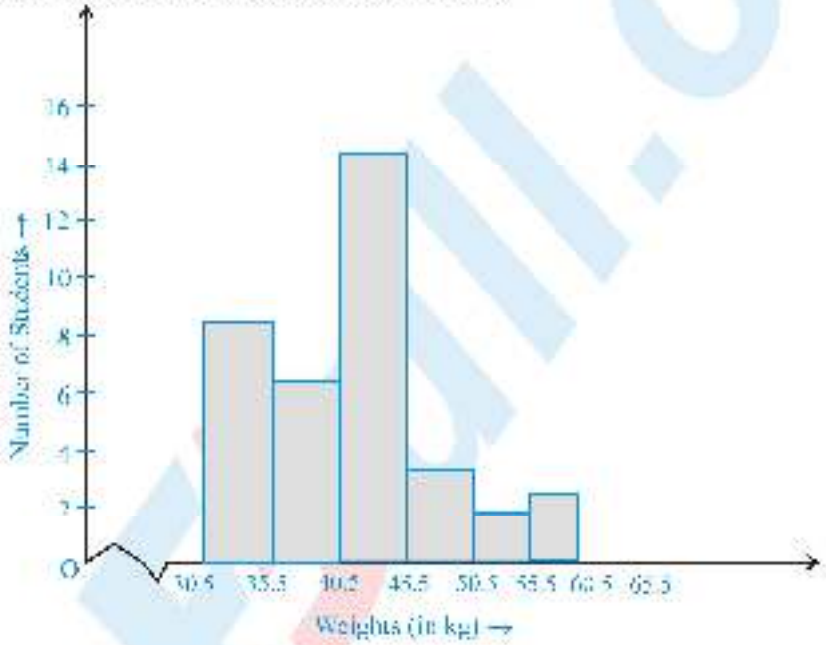
The maximum number of students were born in the month of August.

(B) Histogram

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

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

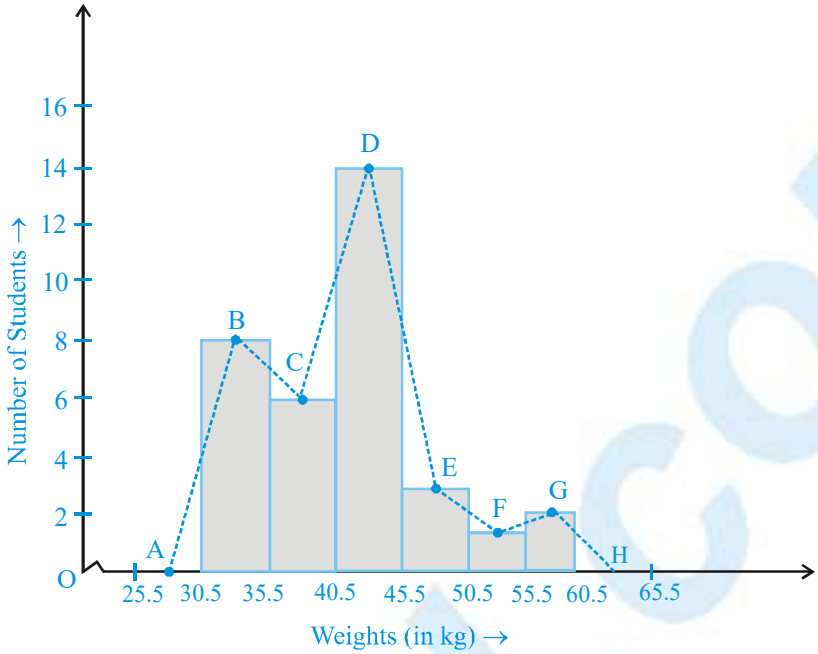
Let us represent the data given above graphically as follows :



Since there are no gaps in between consecutive rectangles, the resultant graph appears like a solid figure. This is called a histogram. Unlike a bar graph, the width of the bar plays a significant role in its construction. Here, in fact areas of the rectangles erected are proportional to the corresponding frequencies since the widths of the rectangles are all equal.

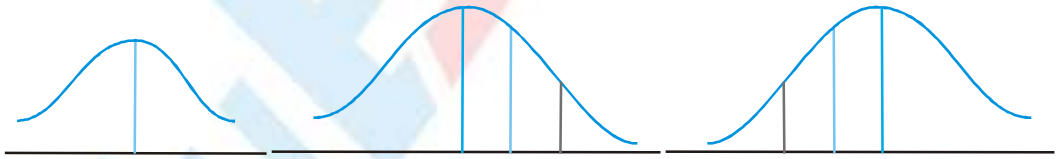
(C) Frequency Polygon

There is yet another visual way of representing quantitative data and its frequencies. Consider the histogram represented by figure shown above. Let us join the mid-points of the upper sides of the adjacent rectangles of this histogram by means of line segments. Let us call these mid-points B, C, D, E, F and G. When joined by line segments, we obtain the figure BCDEFG. 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 given frequency polygon corresponding to the given data.



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. Frequency polygons can also be drawn independently without drawing histogram. Frequency polygons are used when the data is continuous 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.

Skewness : We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. The term 'skewness' refers to lack of symmetry. We can define skewness of a distribution as the tendency of a distribution to depart from symmetry. In a symmetrical distribution we have Mean = Median – Mode. When the distribution is not symmetrical, it is called asymmetrical or skewed. In a skewed distribution Mean ≠ Median ≠ Mode. In positively skewed distribution we have Mean > Median > Mode. In Negatively skewed distribution, we have Mean < Median < Mode.



MEASURES OF DISPERSION

The dispersion of a statistical distribution is the measure of deviation of its values about the their average (central) value.

It gives an idea of scatteredness of different values from the average value.

Generally the following measures of dispersion are commonly used.

(i) Range **(ii) Mean Deviation** **(iii) Variance and Standard Deviation**

(i) Range

The difference between the greatest and least values of variate of a distribution, are called the range of that distribution.

If the distribution is grouped distribution, then its range is the difference between upper limit of the maximum class and lower limit of the minimum class.

Also, coefficient of range = $\frac{\text{difference of extreme values}}{\text{sum of extreme values}}$

Ex. Find the range of following numbers 10, 8, 12, 11, 14, 9, 6

Sol. Here greatest value and least value of the distribution are 14 and 6 resp. therefore
Range = 14 – 6 = 8

Ex. Calculate mean deviation about median for the following data 3, 9, 5, 3, 12, 10, 18, 4, 7, 19, 21.

Sol. Data in ascending order is 3, 3, 4, 5, 7, 9, 10, 12, 18, 19, 21

$$\text{Median} = \frac{n+1}{2}\text{ th value} = 6^{\text{th}}\text{ value} = 9$$

$$\text{Mean deviation about median} = \frac{\sum_{i=1}^{11}|x_i - \text{median}|}{11} = \frac{58}{11}$$

(ii) Mean Deviation

Mean deviation about a central value is defined as the arithmetic mean of the absolute deviations of all the values taken about that central value.

(A) Mean Deviation of Individual Distribution : If $x_1, x_2, ..., x_n$ are n values of a variable x, then the mean deviation from an average A is given by

$$\text{M.D.(A)} = \frac{1}{n}\sum_{i=1}^{n}|x_i - A| = \frac{1}{n}\Sigma|d_i|, \text{ where } d_i = x_i - A$$

(B) Mean Deviation of Discrete Frequency Distribution : If $x_1, x_2, ..., x_n$ are n observation with frequencies $f_1, f_2, ..., f_n$, then mean deviation from an average A is given by -

$$\text{M.D. (A)} = \frac{1}{N}\Sigma f_i|x_i - A| \qquad \text{where } N = \sum_{i=1}^{n} f_i$$

(C) Mean Deviation of Continuous Frequency Distribution : For calculating mean deviation of a continuous frequency distribution, the procedure is same as for a discrete frequency distribution. The only difference is that here we have to obtain the midpoints of the various classes and take the deviations of these mid point from the given average A.

❖ $\text{Coefficient of Mean Deviation} = \frac{\text{Mean Deviation}}{\text{Arithmetic Mean}} = \frac{\text{M.D.}}{\bar{x}}$

❖ Mean deviation of a given set of observations is least when taken about their median.

Ex. Find the mean deviation of numbers 3, 4, 5, 6, 7

Sol. Here $n = 5$, $\bar{x} = 5$

$\therefore$ Mean deviation $= \frac{\sum |x_i - \bar{x}|}{n}$

$= \frac{1}{5}[|3-5| + |4-5| + |5-5| + |6-5| + |7-5|]$

$= \frac{1}{5}[2+1+0+1+2] = \frac{6}{5} = 1.2$

Ex. Find mean deviation from mean

x	5	7	9	10	12	15
f	8	6	2	2	2	6

Sol.

x	f	fx	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$	$f\lvert x - \bar{x} \rvert$
5	8	40	-4	4	32
7	6	42	-2	2	12
9	2	18	0	0	0
10	2	20	1	1	2
12	2	24	3	3	6
15	6	90	6	6	36
	$N = 26$	$\sum fx = 234$			$\sum f\lvert x - \bar{x} \rvert = 88$

$\bar{x} = \frac{\sum fx}{\sum f} = 9$ Now, M.D. $(\bar{x}) = \frac{\sum f|x - \bar{x}|}{\sum f} = \frac{88}{26} = 3.38$

Ex. Find the mean deviation about the median of the following frequency distribution :

Class	0 - 6	6 - 12	12 - 18	18 - 24	24 - 30
Frequency	8	10	12	9	5

Sol. Calculation of mean deviation about the median

Class	Mid values (x_i)	Frequency (f_i)	Cumulative Frequency (c.f.)	$\lvert x_i - M \rvert$	$f_i \lvert x_i - M \rvert$
0 - 6	3	8	8	11	88
6 - 12	9	10	18	5	50
12 -18	15	12	30	1	12
18- 24	21	9	39	7	63
24 -30	27	5	44	13	65
			$N = \sum f_i = 44$		$\sum f_i \lvert x_i - 14 \rvert = 278$

Here $N = 44$, so $\frac{N}{2} = 22$ and the cumulative frequency just greater than $\frac{N}{2}$ is 30. Thus 12-18 is the median class.

Now Median $= l + \frac{N/2 - F}{f} \times h$, where $l = 12, h = 6, f = 12, F = 18$

or Median $= 12 + \frac{22 - 18}{12} \times 6 = 12 + \frac{24}{12} = 14$

Mean deviation about median $= \frac{1}{N}\sum f_i |x_i - 14| = \frac{278}{44} = 6.318$

Limitations of Mean Deviation

Following are some limitations or demerits of mean deviation.

(i) In a frequency distribution the sum of absolute values of deviations from the mean is always more than the sum of the deviations from median. Therefore, mean deviation about mean is not very scientific Thus, in many cases, mean deviation may give unsatisfactory results.

(ii) In a distribution, where the degree of variability is very high, the median is not a representative central value. Thus, the mean deviation about median calculated for such series can not be fully relied.

(iii) In the computation of mean deviation we use absolute values of deviations. Therefore, it cannot be subjected to further algebraic treatment.

VARIANCE AND STANDARD DEVIATION

The variance of a variate x is the arithmetic mean of the squares of all deviations of x from the arithmetic mean of the observations and is denoted by var(x) or σ^2 .

The positive square root of the variance of a variate x is known as standard deviation

i.e. standard deviation (S.D.) = $\sqrt{\text{var}(x)} = \sqrt{\sigma^2} = \sigma$

(i) Variance of Ungrouped Distribution

If $x_1, x_2, \ldots\ldots, x_n$ are n values of a variable x, then by definition

$$\text{var}(x) = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2$$

If the values of variable x are large, the calculation of variance from the above formulae is quite tedious and time consuming. In that case, we take deviation from an arbitrary point A (say) then

$$\text{var}(x) = \frac{1}{n}\left[\sum_{i=1}^{n}(d_i - \bar{d})^2\right] = \frac{1}{n}\sum_{i=1}^{n}d_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}d_i\right)^2, \quad \text{where} \quad d_i = x_i - A$$

$$\text{Also var}(x) = \frac{h^2}{n}\left[\sum_{i=1}^{n}(u_i - \bar{u})^2\right] = h^2\left[\frac{1}{n}\sum_{i=1}^{n}u_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}u_i\right)^2\right] \quad \text{where} \quad u_i = \frac{x_i - A}{h}$$

(ii) Variance of Discrete Frequency Distribution

If $x_1, x_2, \ldots, x_n$ are n observation with frequencies $f_1, f_2, \ldots, f_n$, then

$$\text{var}(x) = \frac{1}{N}\left\{\sum_{i=1}^{n}f_i(x_i - \bar{x})^2\right\} = \left(\frac{1}{N}\sum_{i=1}^{n}f_i x_i^2\right) - \bar{x}^2 \qquad \text{where } N = \sum_{i=1}^{n}f_i$$

If the value of x or f are large, we take the deviations of the values of variable x from an arbitrary point A. (say)

$\therefore$ $d_i = x_i - A$; $i = 1, 2, \ldots, N$

$\therefore$ $$\text{Var}(x) = \frac{1}{N}\left[\sum_{i=1}^{N}f_i(d_i - \bar{d})^2\right] = \frac{1}{N}\left(\sum_{i=1}^{n}f_i d_i^2\right) - \left(\frac{1}{N}\sum_{i=1}^{n}f_i d_i\right)^2 \qquad \text{where } N = \sum_{i=1}^{n}f_i$$

Sometimes $d_i = x_i - A$ are divisible by a common number h(say)

then $$u_i = \frac{x_i - A}{h}, \; i = 1, 2, \ldots\ldots, N$$

then $$\text{var}(x) = \frac{h^2}{N}\left[\sum_{i=1}^{N}f_i(u_i - \bar{u})^2\right] = h^2\left[\frac{1}{N}\sum_{i=1}^{n}f_i u_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n}f_i u_i\right)^2\right] \qquad \text{where } N = \sum_{i=1}^{n}f_i$$

(iii) Variance of a Grouped or Continuous Frequency Distribution

In a grouped or continuous frequency distribution, any of the formulae discussed in discrete frequency distribution can be used.

❖ Coefficient of standard deviation $= \frac{\text{S. D.}}{\text{Mean}} = \frac{\sigma}{\bar{x}}$

(i) Variance is independent of change of origin but dependent on change of scale. Adding or subtracting a positive number from each observation of a group does not affect the variance. If each observation is multiplied by a constant h then variance of the resulting group becomes h^2 times the original variance.

(ii) While calculating S.D., the deviations are to be taken about arithmetic mean only.

COEFFICIENT OF VARIATION

The mean deviation and standard deviation have the same units in which the data is given. Whenever we want to compare the variability of two series with data expressed in different units, we require measure of dispersion which is independent of the units. This measure is coefficient of variation (C.V.)

$$\text{C.V.} = \frac{\text{S.D}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

The series having greater C.V. is said to be more variable and less consistent than the other.

MATHEMATICAL PROPERTIES OF VARIANCE

(i) $\text{Var.}(x_i + \lambda) = \text{Var.}(x_i)$

$\text{Var.}(\lambda x_i) = \lambda^2 . \text{Var}(x_i)$

$\text{Var}(ax_i + b) = a^2 . \text{Var}(x_i)$

where λ, a, b, are constant

(ii) If means of two series containing n_1, n_2 terms are $\bar{x}_1$, $\bar{x}_2$ and their variance's are σ_1^2, σ_2^2 respectively and their combined mean is $\bar{x}$ then the variance σ^2 of their combined series is given by following formula

$$\sigma^2 = \frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{(n_1 + n_2)} \quad \text{where} \quad d_1 = \bar{x}_1 - \bar{x}, d_2 = \bar{x}_2 - \bar{x}$$

i.e. $$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2$$

STANDARD DEVIATION OF COMBINED GROUP

Let $\bar{x}_1, \bar{x}_2$ are A.M. and σ_1, σ_2 are S.D. of two groups having number of observations as n_1 and n_2 respectively then combined standard deviation σ of all the observations taken together is given by

$$\sigma = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}} \quad \text{where} \quad d_1 = \bar{x}_1 - \bar{x}, d_2 = \bar{x}_2 - \bar{x} \text{ and } \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Ex. Find the mean and variance of first n natural numbers.

Sol. $\bar{x} = \frac{\sum x}{n} = \frac{1+2+3+\ldots+n}{n} = \frac{n+1}{2}$

$\text{Variance} = \frac{\sum x^2}{n} - (\bar{x})^2 = \frac{1^2+2^2+3^2+\ldots+n^2}{n} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$

Ex. If $\sum_{i=1}^{18}(x_i - 8) = 9$ and $\sum_{i=1}^{14}(x_i - 8)^2 = 45$, then find the standard deviation of $x_1, x_2, \ldots, x_{18}$

Sol. Let $(x_i - 8) = d_i$

$\therefore \quad \sigma_x = \sigma_d = \sqrt{\frac{\sum d_i^2}{n} - \left(\frac{\sum d_i}{n}\right)^2} = \sqrt{\frac{45}{18} - \left(\frac{9}{18}\right)^2} = \sqrt{\frac{5}{2} - \frac{1}{4}} = \frac{3}{2}$

Ex. Find the variance and standard deviation of the following frequency distribution :

Variable (x_i)	2	4	6	8	10	12	14	16
Frequency (f_i)	4	4	5	15	8	5	4	5

Sol. Calculation of variance and standard deviation

Variable x_i	Frequency (f_i)	$(f_i x_i)$	$x_i - \bar{X}$ $= x_i - 9$	$(x_i - \bar{X})^2$	$f_i (x_i - \bar{X})^2$
2	4	8	-7	49	196
4	4	16	-5	25	100
6	5	30	-3	9	45
8	15	120	-1	1	15
10	8	80	1	1	8
12	5	60	3	9	45
14	4	56	5	25	100
16	5	80	7	49	245
	$N = \sum f_i = 50$	$\sum f_i x_i = 450$			$\sum f_i (x_i - \bar{X})^2 = 754$

Here $N = 50, \sum f_i x_i = 450$

$\therefore \quad \bar{X} = \frac{1}{N}(\sum f x) = \frac{450}{50} = 9$

We have $\sum f_i (x_i - \bar{X})^2 = 754$

$\therefore \quad \text{Var}(X) = \frac{1}{N}\sum f_i (x_i - \bar{X})^2 = \frac{754}{50} = 15.08$

$\text{S.D.} = \sqrt{\text{Var}(X)} = \sqrt{15.08} = 3.88$

Ex. Calculate the mean and standard deviation for the following distribution :

Marks	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80	80 - 90
No. of Students	3	6	13	15	14	5	4

Sol. Calculation of Standard deviation

Class interval	Frequency f_i	Mid - values x_i	$u_i = \frac{x_i - 55}{10}$	$f_i u_i$	u_i^2	$f_i u_i^2$
20 - 30	3	25	-3	-9	9	27
30 - 40	6	35	-2	-12	4	24
40 - 50	13	45	-1	-13	1	13
50 - 60	15	55	0	0	0	0
60 - 70	14	65	1	14	1	14
70 - 80	5	75	2	10	4	20
80 - 90	4	85	3	12	9	36
	$N = \sum f_i = 60$			$\sum f_i u_i = 2$		$\sum f_i u_i^2 = 134$

Here $N = 60, \sum f_i u_i = 2, \sum f_i u_i^2 = 134$ and $h = 10$

$\therefore$ Mean $= \bar{X} = A + h\left(\frac{1}{N}\sum f_i u_i\right) \Rightarrow \bar{X} = 500 + 10\left(\frac{1}{60}\right) = 55.333$

$$\text{Var}(X) = h^2\left[\frac{1}{N}\sum f_i u_i^2 - \left(\frac{1}{N}\sum f_i u_i\right)^2\right] = 100\left[\frac{134}{60} - \left(\frac{2}{60}\right)^2\right] = 222.9$$

$\therefore$ S.D. $= \sqrt{\text{Var}(X)} = \sqrt{222.9} = 14.94$

MEAN SQUARE DEVIATION

The mean square deviation of a distribution is the mean of the square of deviations of variate from assumed mean. It is denoted by S^2.

Hence $S^2 = \frac{\Sigma(x_i - a)^2}{n} = \frac{\Sigma d_i^2}{n}$ (for ungrouped dist.)

$S^2 = \frac{\Sigma f_i(x_i - a)^2}{N} = \frac{\Sigma f_i d_i^2}{N}$ (for freq. dist.).

where $d = (x_i - a)$

Ex. The mean square deviation of a set of n observations $x_1, x_2, \ldots, x_n$ about a point c is defined as $\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$. The mean square deviation about 2 and 2 are 18 and 10 respectively. Then find standard deviation of this set of observations.

Sol. $\frac{1}{n}\Sigma(x_i - 2)^2 = 18$ and $\frac{1}{n}\Sigma(x_i - 2)^2 = 10$

$\Rightarrow \Sigma(x_i - 2)^2 = 18n$ and $\Sigma(x_i - 2)^2 = 10n$

$\Rightarrow \Sigma(x_i - 2)^2 + \Sigma(x_i - 2)^2 = 28n$ and $\Sigma(x_i - 2)^2 - \Sigma(x_i - 2)^2 = 8n$

$\Rightarrow 2\Sigma x_i^2 = 8n - 28n$ and $8\Sigma x = 8n$

$\Rightarrow \quad \Sigma x_i^2 = 10n \quad$ and $\quad \Sigma x_i = n$

$\Rightarrow \quad \frac{\Sigma x_i^2}{n} = 10 \quad$ and $\quad \frac{\Sigma x_i}{n} = 1$

$\therefore \quad \sigma = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2} = \sqrt{10 - (1)^2} = 3$

RELATION BETWEEN VARIANCE AND MEAN SQUARE DEVIATION

$\Rightarrow \quad \sigma^2 = \frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2$

$\Rightarrow \quad \sigma^2 = s^2 - d^2, \quad$ where $d = x - a = \frac{\Sigma f d}{N}$

$\Rightarrow \quad s^2 = \sigma^2 + d^2 \quad \Rightarrow \quad s^2 \geq \sigma^2$

Hence the variance is the minimum value of mean square deviation of a distribution

Ex. Determine the variance of the following frequency dist.

class	0 - 2	2 - 4	4 - 6	6 - 8	8 - 10	10 - 12
f_i	3	7	13	19	9	1

Sol. Let $a = 7, h = 2$

class	x_i	f_i	$u_i = \frac{x_i - a}{h}$	$f_i u_i$	$f_i u_i^2$
0 - 2	1	3	-3	-9	27
2 - 4	3	7	-2	-14	28
4 - 6	5	13	-1	-13	13
6 - 8	7	19	0	0	0
8 - 10	9	9	1	9	9
10 - 12	11	1	2	2	4
	$N = 50$			$\Sigma f_i u_i = -21$	$\Sigma f_i u_i^2 = 81$

$\therefore \quad \sigma^2 = h^2\left[\frac{\Sigma f_i u_i^2}{N} - \left(\frac{\Sigma f_i u_i}{N}\right)^2\right] = 4\left[\frac{81}{50} - \left(\frac{21}{50}\right)^2\right] = 4[1.62 - 0.1764] = 4.97$

Statistics is the Science of collection, organization, presentation, analysis and interpretation of the numerical data.

Useful Terms

1. Limit of the Class

The starting and end values of each class are called **Lower** and **Upper limits**.

2. Class Interval

The difference between upper and lower boundary of a class is called class interval or size of the class.

3. Primary and Secondary Data

The data collected by the investigator himself is known as the **primary data**, while the data collected by a person other than the investigator is known as the **secondary data**.

4. Variable or Variate

A characteristics that varies in magnitude from observation to observation. e.g. weight, height, income, age, etc are variables.

5. Frequency

The number of times an observation occurs in the given data, is called the frequency of the observation.

6. Discrete Frequency Distribution

A frequency distribution is called a distance frequency distribution, if data are presented in such a way that exact measurements of the units are clearly shown.

7. Continuous Frequency Distribution

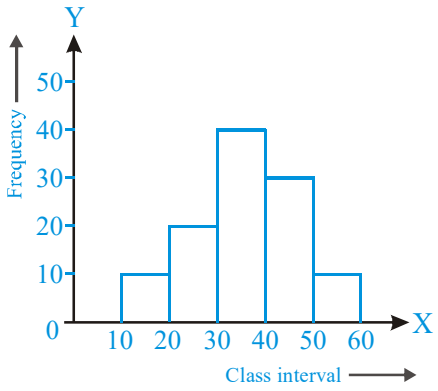
A frequency distribution which data are arranged in classes (or groups) which are not exactly measurable.

8. Cumulative Frequency Distribution

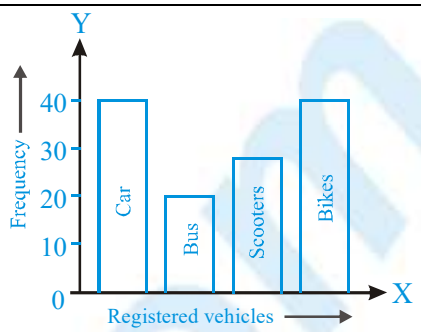
Suppose the frequencies are grouped frequencies or class frequencies. If however, the frequency of the first class is added to that of the second and this sum is added to that of the third and so on, then the frequencies, so obtained are known as cumulative frequencies (cf).

(i) **Histogram:** to draw the histogram of a given continuous frequency distribution, we first mark off all the class intervals along X-axis on a suitable scale. On each of these class intervals on the horizontal axis, we eract (vertical) a rectangle whose height is proportional to the frequency of that particular class, so that the area of the rectangle is proportional to the frequency of the class.

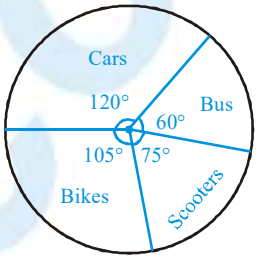
If however the classes are of unequal width, then the height of the rectangles will be proportional to the ratio of the frequencies to the width of the classes.



(ii) **Bar Diagrams:** In bar diagrams, only the length of the bars are taken into consideration. To draw a bar diagram, we first mark equal lengths for the different classes on the horizontal axis, i.e,. X-axis. On each of these lengths on the horizontal axis, we erect (vertical) a rectangle whose heights are proportional to the frequency of the class.



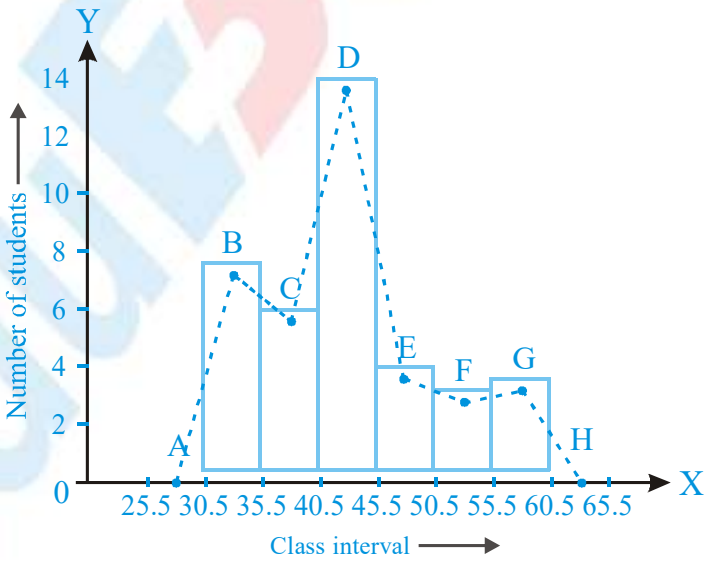
(iii) **Pie Diagrams:** Pie are used to represent a relative frequency distribution. A pie diagram consists of a circle divided into as many sectors as there are classes in a frequency distribution. The area of each sector is proportional to the relative frequency of the class.



Now, we make angles at the centre proportional to the relative frequencies. And in order to get the angles of the desired sectors, we divide 360° in the proportion of the various relative frequencies, i.e,.

$$\text{Central angle} = \left[\frac{\text{Frequence} \times 360°}{\text{Total frequency}}\right]$$

(iv) **Frequency Polygon** To draw the frequency polygon of an ungrouped frequency distribution, we plot the points with abscissas as the variate values and the ordinate as the corresponding frequencies. These plotted points are joined by straight lines to obtain the frequency polygon.

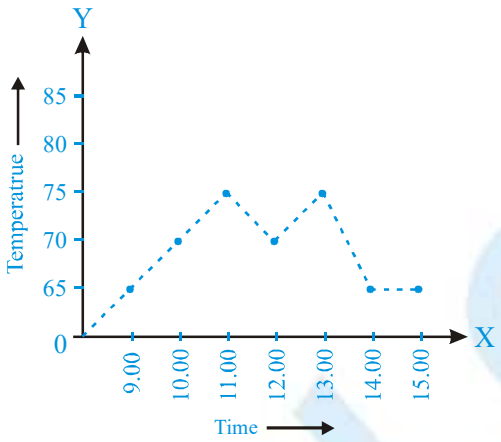


(v) Cumulative Frequency Curve (Ogive): The term ogive is pronounced as ogive. It is a shape consisting of a concave arc flowing into a convex arcs. i.e., forming as S-shaped curve with vertical ends.

There are two methods of constructing an ogiven,

(i) 'less than' type ogive

(ii) 'more than' type ogive.



Measures of Central Tendency

Generally, average value of a distribution in the middle part of the distribution, such type of values are known as measures of central tendency.

The following are the five measures of central tendency :

1. Arithmetic Mean
2. Geometric Mean
3. Harmonic Mean
4. Median
5. Mode

1. Arithmetic Mean

The arithmetic (or simple) mean of a set of observations is obtained by dividing the sum of the values of observations by the number of observations.

(i) Arithmetic Mean for Unclassified (Ungrouped) Data If n observations, $x_1, x_2, x_3, \ldots, x_n$, then their arithmetic mean

$$A \text{ or } \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(ii) Arithmetic Mean for Frequency Distribution

Let $f_1, f_2, \ldots, f_n$ be corresponding frequencies of $x_1, x_2, \ldots, x_n$.

Then, arithmetic mean

$$A = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

(iii) Arithmetic Mean for Classified (Grouped) Date

For a classified data, we take the class marks $x_1, x_2, ..., x_n$ of the classes as variables, then arithmetic mean by

(A) Direct method is $A = \dfrac{\sum_{i=1}^{n} xf}{\sum_{i=1}^{n} f}$

(B) Deviation Method is $A = A_1 + \left(\dfrac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}\right) h$

(C) Step Deviation method is $\bar{x} = A_1 + \dfrac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i} \times h$

where, A_1 = assumed mean

u_i = step deviation $= \dfrac{x_i - A_i}{h}$ and h = width of interval.

(iv) Combined Mean: If $A_1, A_2, ... A_n$ are the n arithmetic mean having number of corresponding observations $n_1, n_2, ... n_r$, then arithmetic mean of the combined group x is called the combined mean of the observation

$$A = \frac{n_1A_1 + n_2A_2 + ... + n_rA_r}{n_1 + n_2 + ... + n_r} = \frac{\sum_{i=1}^{n} n_iA_i}{\sum_{i=1}^{n} n_if_i}$$

(v) Weighted Arithmetic Mean If $w_1, w_2, ..., w_n$ are the weights assigned to the values $x_1, x_2, ... x_n$ respectively, then the weights assigned to the values $x_1, x_2, ..., x_n$ respectively, then the weighted arithmetic mean is

$$A_w = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i}$$

Properties of Arithmetic Mean

(i) Mean is dependent of change of origin and changes of scale.

(ii) Algebraic sum of the deviations of a set of values from their arithmetic mean is zero.

(iii) The sum of the squares of the deviations of a set of values is minimum when taken about mean.

2. Geometric Mean

(i) If $x_1, x_2, ..., x_n$ be n non-zero observations, then their geometric mean is defined as

$$G = \sqrt[n]{x_1x_2...x_n}$$

or $$G = \text{antilog}\left[\frac{\log x_1 + \log x_2 + ... + \log x_n}{n}\right]$$

(ii) Let $f_1, f_2, ..., f_n$ be the corresponding frequencies of non-zero observations $x_1, x_2, ..., x_n$, then geometric mean is defined as :

$$G = \left(x_1^{f_1}x_2^{f_2}....x_n^{f_n}\right)^{\frac{1}{N}}, \quad \text{where} \quad N = \sum_{i=1}^{n} f_i$$

or $$G = \left[\frac{1}{N}(f_i \log x_1 + f_2 \log x_2 + ... + f_n \log x_n\right]$$

3. Harmonic Mean (HM)

The harmonic mean of n non-zero observations $x_1, x_2, ..., x_n$ is defined as

$$HM = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+...+\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

It their corresponding frequencies $f_1, f_2, ..., f_n$ respectively, then

$$HM = \frac{f_1+f_2+....+f_n}{\left(\frac{f_1}{x_1}+\frac{f_2}{x_2}+...\frac{f_n}{x_n}\right)} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n}\frac{f_i}{x_i}}$$

4. Median

The median of a distribution is the value of the middle variable, when the variables are arranged in ascending or descending order.

Median (M_d) is an average of position of the numbers.

(i) Median for Simple Distribution

Firstly, arrange the data in ascending or descending order and then find the number of observations n.

(A) If n is odd, then $\left(\frac{n+1}{2}\right)$ th term is the median.

(B) If n is even, then there are two middle terms namely $\left(\frac{n}{2}\right)$ th and $\left(\frac{n}{2}+1\right)$ th terms.

Hence,

$$\text{Median} = \text{Mean of } \left(\frac{n}{2}\right) \text{th and } \left(\frac{n}{2}+1\right) \text{th observations}$$

$$= \frac{1}{2}\left[\left(\frac{n}{2}\right)\text{th} + \left(\frac{n}{2}+1\right)\text{th}\right] \text{ of observations}$$

(ii) Median for Unclassified (Ungrouped) Frequency Distribution

(i) Firstly, find $\frac{N}{2}$, where $N = \sum_{i=1}^{n} f_i$.

(ii) Find the cumulative frequency of each value of the variable and take value of the variable which is equal to or just greater than $\frac{N}{2}$.

(iii) This value of the variable is the required median.

(iii) Median for Classified (Grouped) Data

If in a continuous distribution, the total frequency be N, then the class whose cumulative frequency is either equal to N/2 or is just greater than N/2 is called **median class**.

For a continuous distribution, median

$$M_d = l + \frac{\frac{N}{2} - C}{f} \times h$$

where, l = lower limit of the median class

f = frequency of the median class

N = total frequency = $\sum_{i=1}^{n} f_i$

C = cumulative frequency of the class just before the median class

h = length of the median class

Note

The intersection point of less than ogive and more than ogive is the median.

5. Mode

The mode (M_0) of a distribution if the value of the point about which the observations tend to be most heavily concentrated. It is generally the value of the variable which appears to occur most frequently in the distribution.

(i) Mode for a Simple Data

The value which is repeated maximum number of times, is the required mode.

e.g. Mode of the data 70, 80, 90, 96, 70, 96, 96, 90 is 96 as 96 occurs maximum number of times.

(ii) Mode for Unclassified (Ungrouped) Frequency Distribution

Mode is the value of the variate corresponding to the maximum frequency.

(iii) For Classified (Group) Distribution

The class having the maximum frequency is called the **modal class** and the middle point of the model class is called the **crude mode**. The class just before the modal class is called pre-modal class and the class after the modal class is called the post-modal class.

Mode for classified data (Continuous Distribution) is given by

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

where, l = lower limit of the modal class

f_0 = frequency of the modal class

f_1 = frequency of the pre-modal class

f_2 = frequency of the post-modal class

h = length of the class interval

Relation between Mean, Median and Mode

(i) Mean – Mode = 3(Mean – Median)

(ii) Mode = 3 Median – 2 Mean

Measure of Dispersion

The degree to which numerical data tend to spread about an average value is called the dispersion of the data. The four measure of dispersion are

1. Range **2.** Mean deviation

3. Standard deviation **4.** Root mean square deviation

1. Range

The difference between the highest and the lowest observation of a data is called its range.

i.e., Range $X_{max} - X_{min}$

$\therefore$ The coefficient of range $= \dfrac{X_{max} - X_{min}}{X_{max} + X_{min}}$

It is widely used in statistical series relating to quality control in production.

(i) Inter quartile range $= Q_3 - Q_1$

(ii) Semi-inter quartile range (Quartile deviation) and coefficient of quartile deviation $= \dfrac{Q_3 - Q_1}{Q_3 + Q_1}$

3. Mean Deviation (MD)

The arithmetic mean of the absolute deviations of the values of the variable from a measure of their average (mean, median, mode) is called Mean Deviation (MD). It is denoted by δ.

(i) For simple (discrete) distribution

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

where, n = number of terms, $\bar{x}$ = A or M_d or M_0

(ii) For unclassified frequency distribution $\delta = \dfrac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{\sum_{i=1}^{n} f_i}$

(iii) For classify d distribution $\delta = \dfrac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{\sum_{i=1}^{n} f_i}$

where x_i is the class mark of the interval.

Note The mean deviation is the least when measured from the median.

Coefficient of Mean Deviation

It is the ratio of MD and the mean from which the deviation is measured.

Thus, the coefficient of MD $= \dfrac{\delta_A}{A}$ or $\dfrac{\delta_{M_d}}{M_d}$ or $\dfrac{\delta_{M_O}}{M_O}$

Limitations of Mean Deviation

(i) If the data is more scattered or the degree of variability is very high, then the median is not a valid representative.

(ii) The sum of the deviations from the mean is more than the sum of the deviations from the median.

(iii) The mean deviation is calculated on the basis of absolute values of the deviations and so cannot be subjected to further algebraic treatment.

3. Standard Deviation and Variance

Standard deviation is the square root of the arithmetic mean of the squares of deviations of the terms from their AM and it is denoted by σ.

The square of standard deviation is called the **variance** and it is denoted by the symbol σ^2.

(i) For simple distribution

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \frac{1}{n}\sqrt{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}$$

where, n is a number of observations and $\bar{x}$ is mean.

(ii) For discrete frequency distribution

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}f(x_i - \bar{x})^2}{N}} = \frac{1}{N}\sqrt{N\sum_{i=1}^{n}f_i x_i^2 - \left(\sum_{i=1}^{n}f_i x_i\right)^2}$$

Shortcut Method

$$\sigma = \frac{1}{N}\sqrt{N\sum_{i=1}^{n}f_i d_i^2 - \left(\sum_{i=1}^{n}f_i d_i\right)^2}$$

where, d_i = deviation from assumed mean = $x_i - A$ and A = assumed mean

(iii) For continuous frequency distribution

$$\sigma = \sqrt{\frac{\sum f_i(x_i - \bar{x})^2}{N}}$$

where, x_i is class mark of the interval.

Shortcut Method

$$\sigma = hN\sqrt{\sum_{i=1}^{n}f_i u_i^2 - \left(\sum_{i=1}^{n}f_i u_i\right)^2}$$

where, $u_i = \frac{x_i - A}{h}$, A = assumed mean and h = width of the class

Standard Deviation of the Combined Series

If n_1, n_2 are the sized, $\bar{X}_1$, $\bar{X}_2$ are the means and σ_1, σ_2 are the standard deviation of the series, then the standard deviation of the combined series is

$$\sigma = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}}$$

where, $d_1 = \bar{X}_1 - \bar{X}$, $d_2 = \bar{X}_2 - \bar{X}_2$

Effects of Average and Dispersion on Change of origin and Scale

	Change of origin	Change of scale
Mean	Dependent	Change of Scale
Median	Not dependent	Dependent
Mode	Not dependent	Dependent
Standard Deviation	Not dependent	Dependent
Variance	Not dependent	Dependent

Note

(i) Change origin means either subtract or add in observations.

(ii) Change of scale means either multiply or divide in observations.

Important Point to be Remember

(i) The ratio of SD(σ) and the AM ($\bar{x}$) is called the coefficient of standard deviation $\left[\frac{\sigma}{\bar{x}}\right]$.

(ii) The percentage from of coefficient of SD i.e., $\left(\frac{\sigma}{\bar{x}}\right) \times 100$ is called coefficient of variation.

(iii) The distribution for which the coefficient of variation is less is called more consistent.

(iv) Standard deviation of First n natural number is $\sqrt{\frac{n^2-1}{12}}$.

(v) Standard deviation is independent of change of origin, but it is depend on change of scale.

(vi) Quartile deviation $= \frac{2}{3}$ Standard deviation.

(vii) Mean deviation $= \frac{4}{5}$ standard deviation.

4. Root Mean Square Deviation (RMS)

The square root of the AM of squares of the deviations from an assumed mean is called the root mean square deviation.

Thus,

(i) For simple (discrete) distribution

$$S = \sqrt{\frac{\sum (x - A')^2}{n}}, \text{ where } A' = \text{assumed mean}$$

(ii) For frequency distribution

$$S = \sqrt{\frac{\sum f(x - A')^2}{\sum f}}$$

If $A' = A$ (mean), then $S = \sigma$

Important Points to be Remember

(i) The RMS deviation is the least when measured from AM.

(ii) The sum of the squares of the deviation of the values of the variable is the least when measured from AM.

(iii) $\sigma^2 = A^2 - \frac{\sum f x^2}{\sum f}$

(iv) For discrete distribution, if $f = 1$, then $\sigma^2 = A^2 - \frac{\sum x^2}{n}$

(v) The mean deviation about the mean is less than or equal to the SD i.e., MD $\leq \sigma$.